Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 1849995, 13 pages
https://doi.org/10.1155/2022/1849995

Research Article

An Image Saliency Detection Method Based on Combining Global and Local Information

Hangxu Yang,[1,2] **Yongjian Gong**,[1] **and Kai Wang**[3]

[1]*College of Mechanical and Electrical Engineering, Jinhua Polytechnic, Jinhua 321017, Zhejiang, China*
[2]*Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua 321017, Zhejiang, China*
[3]*College of Engineering, Nanjing Agricultural University, Nanjing 210031, Jiangsu, China*

Correspondence should be addressed to Hangxu Yang; yanghx_83@126.com

Received 27 February 2022; Accepted 30 March 2022; Published 19 April 2022

Academic Editor: Zaoli Yang

In the field of computer vision, image saliency target detection can not only improve the accuracy of image detection but also accelerate the speed of image detection. In order to solve the existing problems of the saliency target detection algorithms at present, such as inconspicuous texture details and incomplete edge contour display, this paper proposes a saliency target detection algorithm integrating multiple information. The algorithm consists of three processes: preprocessing process, multi-information extraction process, and fusion optimization process. The frequency domain features of the image are calculated, the algorithm calculates the frequency domain features of the image, introduces power law transform and feature normalization, improves the frequency domain features of the image, saves the information of the target region, and inhibits the information of the background region. On three public MSRA, SED2, and ECSSD image datasets, the proposed algorithm is compared with other classical algorithms in subjective and objective comparison experiments. Experimental results show that the proposed algorithm can not only accurately and comprehensively extract significant target regions but also retain more texture information and complete edge information while satisfying the human visual experience. All evaluation indexes are significantly better than the comparison algorithm, showing good reliability and adaptability.

1. Introduction

Visual attention is essentially a kind of biological extract of the important information from the complicated external environment and the information of interest. Relative to the important information filtering mechanism, this mechanism can be largely simplified complex external image information and decomposition, and then the next to the important information for further processing [1]. Based on a large number of studies, researchers summarize the mechanism of visual attention, which is systematically divided into two directions: one is the bottom-up mechanism of visual attention driven by data information; the other is the top-down visual attention mechanism driven by task information. The bottom-up visual attention mechanism is based on stimulus information that distinguishes it from other areas in the scene or image [2]. After the 1980s, researchers mainly focused on the bottom-up visual attention mechanism driven by data. In the study of the visual attention mechanism, based on Koch's framework and Treisman's feature integration theory, some features of primary vision are proposed, including the color, direction, and brightness of the image. By extracting these features, the saliency map of the three feature dimensions of color, direction, and brightness can be formed, and the saliency image of different dimensions can be merged using the multi-feature map strategy to form an interest map [3]. The interest map usually contains a series of targets that need attention. A single target is selected through a competitive mechanism, and then the focus of attention is shifted. The data-driven visual attention mechanism model uses a series of operations such as image information sampling, feature extraction, and focus of attention search to find the target worthy of attention in the image, forming a focus of attention detection method with good operability and fast computing capabilities [4]. This mechanism model usually simply superimposes different

characteristics of image information such as brightness, color, and direction, but this superposition method is relatively simple and crude, which is very different from the processing of images by the human visual system. Moreover, the region of interest is an important target and only occupies a small area of the image. The matching operation needs to be processed globally, the matching process is complicated, and it is easy to cause computational waste [5, 6].

The top-down visual attention mechanism is driven by task information. According to the given prior information of the specified task, a visual expectation is set in advance, the desired specific target object is derived from the image, and the desired target object in the image is separated, the irrelevant area in the image is removed based on the task information, and the area of interest is filtered out, and then the area of interest in the image can be further processed separately [7]. Under normal circumstances, an image is divided into several visual areas with different priorities, and the desired visual target is included in the image areas with higher visual priority in these different areas. In this way, the image is subject to a regional priority division mechanism. It is similar to the mechanism when humans observe things [8, 9]. The top-down visual attention mechanism is dominated by people's many factors, such as perception, consciousness, and choice, and has passed a subjective choice made by subjective consciousness driven by the target object [10]. The model's focus on achieving different goals is mainly achieved through three aspects: the characteristics of the target object in the scene or image, the prior information of the scene, and the task's requirements for the object. The characteristics of the object refer to the fact that the primary characteristics of the object, including color, brightness, and direction, are not added to the visual attention mechanism, but the characteristics of distinguishing the object from other areas in the image are added [11].

Linear filtering, anisotropic suppression, and statistically effective decision-making are the three defined stages in the top-down visual attention mechanism. Based on these three stages in the top-down visual attention mechanism model, Navalpakkam et al. proposed a visual attention model, guided by task information as the model operation, and established a structure related to the task of the attention position in the scene or image. The attention model consists of four different functional modules: visual brain, proxy, working memory, and long-term memory. The visual brain is divided into three parts, which are composed of the saliency map, the related task map, and the attention guide map, respectively [12]. The visual brain is responsible for generating salient images based on a top-down model of visual attention. The responsibility function of the task correlation map is to make some connections between the input image and the actual visual task. Then, the product of the saliency image and the related task image is carried out point-to-point to obtain an attention wizard graph. The visual brain performs further analysis based on the generated attention guide map, and the region with the strongest saliency is the current saliency region [13]. After the visual brain has analyzed the received information, it sends the processed target information to the agency. The agency's responsibility is to receive the processed target information sent by the visual brain and then process the processed target information. The memory module is responsible for receiving the information from the agency. Information resources are connected in this way at the working memory module. In the same way, such a connection is also established between the agent and the long-term memory module. In this way, communication is realized. The correlation of saliency regions is determined according to the target information sent to each other, and then the correlation is calculated. The current saliency image obtained in advance is roughly estimated, and then the relevant task map is constantly updated. The long-term memory module contains the physical objects of various abstract things in the real nature and the connections between these abstract physical objects. In order to realize the function of information transfer among the other three parts, agents are needed to transfer information from them. In this way, the process of information transfer will continue to be repeated until the relevant task graph is reached and all the relevant entity objects in the input image can be determined. Itti and Navalpakkam improve the ability of the original bottom-up visual attention model to detect and identify targets by performing top-down control using a representation of known target entities. Currently, the most representative top-down visual attention models are the Navalpakkam model and Visual Object Detection with a CompUtational Attention System (VOCUS) top-down model, respectively. Navalpakkam et al. wanted to obtain the weight of each feature, adopted the signal-to-noise ratio of the maximized target and background, and then carried out statistics on the experimental results. The VOCUS top-down model method is different. It is through the calculation of the image of each participating training resource that we get to the characteristics of each target and the background, the ratio of gain weight vector; all training image resources geometric average of the weight vector is the final weight vector; another kind of method of the weight values is obtained by this method. In the above two training models, only when the combination of their target information region and the background region is basically the same or roughly similar to the combination of the target information region and background region in the training of image resources, the experimental results will be the ideal experimental results we expect. However, when the combination of the target information region and background region of the tested image resources is very different from that of the training image resources, the experimental results obtained by the above two models will be very different from the expected experimental results. In addition, it is not difficult to see that the Navalpakkam model and VOCUS top-down model, target information region, background region of training image resources, and the combined relationship between them all play an important role in model training. Therefore, when a new test image resource appears, and the target information region and background region of its training image resource and their combination are similar to the previous training image resource set, the experimental results of the above model will be very close to the expected

experimental results. To sum up, the advantage of the data-driven bottom-up visual attention model is that it has a wide range of applications but is weak in pertinence. When the task image is very specific, the processing result will not be ideal. However, the top-down visual attention model driven by tasks has strong pertinence, but its scope of application is very narrow. When the character information is not clear, the image data can be processed at a loss [14, 15].

The main research of this paper is as follows.

(1) Two models related to image saliency are analyzed, namely, the bottom-up visual attention mechanism model driven by data information and the top-down visual attention mechanism model driven by task.

(2) Analyze several image saliency detection algorithms, including Itti, GBVS algorithm based on graph theory, CA algorithm based on context, and HC based on histogram contrast, etc., and analyze the principles of these four algorithms.

(3) In view of the problems existing in saliency target detection algorithms at the present stage, such as inconspicuous texture detail information description and incomplete edge contour display, a saliency target detection research algorithm integrating multiple information is proposed. The algorithm consists of three processes: preprocessing process, multi-information extraction process, and fusion optimization process.

(4) Test the proposed algorithm with Itti, FT, GBVS, CA, LC, and other visual saliency detection algorithms through experiments, observe the effect of each algorithm running on the data set, and compare the effect of each algorithm through the PR curve.

2. Detection of Image Saliency Region

2.1. Itti Algorithm. Itti model is a selective visual attention model driven by data information and based on saliency. The saliency value represents the contrast ratio of pixels in multiple dimensions, such as color, brightness, and direction, with the surrounding background. The saliency algorithm can be divided into two parts: feature region extraction and saliency map generation [14]. Itti flow is as follows:

(1) Part one: Extraction of feature area.

The first step is to build the Gauss Pyramid.

Firstly, the image needs to build a Gaussian pyramid with the number of layers 9, and a Gaussian linear filter is used here. The Gauss pyramid is made up of three parts, namely brightness, color, and direction [16].

Gaussian downsampling is performed on r, g, and b channels to obtain three-channel images $\gamma(\sigma)$, $g(\sigma)$, and $b(\sigma)$ at nine scales, among which $\sigma \in \{0 \dots 8\}$ is obtained.

Then we started to build a high Gaussian pyramid and calculated $I = (r + g + b)/3$ at each of the 9 scales to get $I(\sigma)$. Then, according to $I(\sigma)$, $\gamma(\sigma)$, $g(\sigma)$, and $b(\sigma)$ are normalized to separate the hue from the brightness. The reason for this is that the hue will be difficult to distinguish in the case of low brightness. The normalization of each pixel point is performed only for the point with brightness $I >$ Maxinum/10, while the remaining points will be zeroed, where Maxinum represents the maximum brightness value in the scale image where the point is located.

Next, we started to build the Gauss pyramid, which was calculated at 9 different scales:

$$\begin{aligned} R(\sigma) &= r(\sigma) - \frac{(g(\sigma) + b(\sigma))}{2}, \\ G(\sigma) &= g(\sigma) - \frac{(r(\sigma) + b(\sigma))}{2}, \\ B(\sigma) &= b(\sigma) - \frac{(r(\sigma) + g(\sigma))}{2}, \\ Y(\sigma) &= \frac{(r(\sigma) + g(\sigma))}{2} - \frac{|r(\sigma) - g(\sigma)|}{2} - b(\sigma). \end{aligned} \tag{1}$$

The above four variables represent the Gauss pyramid with four different colors of red, green, blue, and yellow, respectively.

Finally, the Gabor filter is used to construct a Gabor direction pyramid $O(\sigma, \theta)$ where $\sigma \in \{0 \dots 8\}$ and $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

The second step is to construct the feature map.

Gaussian pyramids of three different dimensions are obtained above, including brightness Gaussian pyramid, color Gaussian pyramid, and directional Gaussian pyramid. After obtaining the Gaussian pyramids of these three dimensions, the center-surround method (Center(c) refers to fine scale, Surround(s) refers to coarse scale). The calculation method is as follows:

$$\begin{aligned} I(c, s) &= |I(c) \Theta I(s)|, \\ RG(c, s) &= |(R(c) - G(c)) \Theta (G(s) - R(s))|, \\ BY(c, s) &= |(B(c) - Y(c)) \Theta (Y(s) - B(s))|, \\ O(c, s, \theta) &= |O(c, \theta) \Theta O(s, \theta)|, \end{aligned} \tag{2}$$

where $c\{2, 3, 4\}$, $s = c + \delta, \delta \in \{3, 4\}$. The above Θ refers to the matrix subtraction operation after adjusting the size of the two images to the same size. I represents the luminance feature map, RG and BY both represent the color feature map, which makes use of the "color positions" system of the cerebral cortex, and O represents the direction feature map. Therefore, a total of $6 + 12 + 6 \times 4 = 42$ feature maps are generated.

(2) Part two: the generation of saliency graph.

The feature graph constructed in Part 1 will be further operated and normalized. The purpose of this operation is to eliminate the amplitude differences associated with features. In order to eliminate noise interference and to realize the goal of making a few salient points with the most prominent significance evenly distributed on the whole feature graph, only a few salient points are reserved on the feature graph for optimization iteration and $N(\cdot)$ is used as the normalization and iteration process. After the normalization and iteration process, multiple different types of saliency graphs can be

obtained, and each saliency graph corresponds to a class of features. Then, the significance graph of the input image can be obtained by averaging these obtained saliency graphs. As an early salient region extraction algorithm, this algorithm can effectively extract salient regions, but due to the use of the adjacent interpolation method in the cross-scale calculation, the resolution of the final salient map is relatively low, and the information of the image edge is lost [17, 18].

2.2. Visual Saliency Detection Algorithm Based on Graph Theory. GBVS is a visual saliency detection algorithm based on graph theory. Since it is an improved Itti algorithm, it is similar to Itti in some operating parts, such as feature extraction. Of course, there are also differences. In the calculation of significance value, it is different from Itti but uses the Markov chain algorithm [19, 20].

The algorithm is divided into three steps:

Step 1: Feature extraction.

Feature vectors are extracted from image plane positions to generate feature maps, which is similar to Itti.

Step 2: Use the Markov chain to generate a saliency map.

First, set the feature diagram of the input image as follows:

$$M\colon [n]^2 \longrightarrow R. \tag{3}$$

Define the distance from pixel points (i, j) to (p, q) in the image as follows:

$$d((i,j)\|(p,q)) = \left|\log\frac{M(i,j)}{M(p,q)}\right|. \tag{4}$$

In some experiments can also use $|M(i,j) - M(i,j)|$ instead.

All points in the feature graph M generated by the input image are connected in pairs to obtain a new image G_A. The weight of directed edges from points (i, j) to (p, q) is defined as follows:

$$w_1((i,j),(p,q)) = d((i,j)\|(p,q)) \times F(i-p, j-q),$$
$$F(a,b) = e^{a^2+b^2/2\sigma^2}. \tag{5}$$

Then normalize the weight of the directed edge emitted by each node, define the Markov chain on G_A, and obtain the significance value from the comparison of nodes in pairs, so as to generate the saliency graph, which is defined as follows:

$$A\colon [n]^2 \longrightarrow R. \tag{6}$$

Step 3: Normalized saliency map.

The purpose of this step is to extract concentrated blocks on the saliency graph by calculating the concentration of blocks. To define the concentration of blocks, the authors propose another Markov chain. For the interconnected points (i, j) and (p, q), including (i, j), the weight of this directed edge is defined as follows:

$$w_2((i,j),(p,q)) = A(p,q)F(i-p, j-q). \tag{7}$$

For consistency, the weights of the edges starting from each point are normalized. According to the Algorithm of the Markov chain, the balanced distribution of the Markov chain is calculated on the nodes. At this time, the blocks will be concentrated towards those nodes with high weights, and the extracted concentrated blocks are the final saliency graph [21, 22].

2.3. Context-Based Visual Saliency Detection Algorithm. The algorithm no longer uses the commonly used RGB color space but uses LAB color space, where L is brightness, A is the value between red and green, and B is the value between yellow and blue. LAB has a wider color gamut and is closer to the perception of color by human eyes. When calculating the difference between image blocks, only the difference between this block and the nearest K blocks is considered, and all blocks in the image are not considered. It is divided into the following three steps:

Step 1: Define the inconsistencies of p_i and p_k in the two regions:

$$d(p_i, p_k) = \frac{d_{\text{color}}(p_i, p_k)}{1 + c^* d_{\text{position}}(p_i, p_k)}, \tag{8}$$

where $d_{\text{color}}(p_i, p_k)$ is the color distance between two blocks, and $d_{\text{position}}(p_i, p_k)$ is the spatial distance between two blocks.

Step 2: Only the K blocks most similar to the candidate blocks are considered, with M scales, then the saliency under a single scale r is as follows:

$$S_i^r = 1 - \exp\left(-\frac{1}{K}\sum_{k=1}^{K} d(p_i^r, p_k^r)\right). \tag{9}$$

Step 3: cross-scale fusion, saliency S is represented by the average saliency of each scale:

$$\overline{S_l} = \frac{1}{M}\sum_{r=1}^{M} S_i^r. \tag{10}$$

In the context based visual saliency detection algorithm, multiscale correction and context perception are combined so that the saliency of the target area in the saliency map is the highest, but the more it spreads to the edge, the lower the saliency, and at the edge, the image saliency is the most blurred [23, 24].

2.4. Visual Saliency Detection Algorithm Based on Histogram Contrast. This algorithm statistics the color features of the input image and then statistics the histogram of the color features for comparison. In this algorithm, the saliency of a

pixel is represented by its color distance from other pixels in the image. In other words, the saliency $S(I_k)$ of pixel I_k in image I is defined as follows:

$$S(I_k) = \sum_{\forall I_i \in I} D(I_k, I_i), \tag{11}$$

where $D(I_k, I_i)$ is the color distance brightness of pixel I_k and pixel I_i in $L*a*b$ space. The formula is expanded in pixel order as follows:

$$S(I_k) = D(I_k, I_1) + D(I_k, I_2) + \cdots + D(I_k, I_N), \tag{12}$$

where N is the number of pixels of image I. Because the spatial relationship in the image is not taken into account, in the above definition, pixels of the same color have the same saliency. Therefore, it is not necessary to calculate the saliency of all pixels in sequence during calculation, but only to calculate the saliency of each color to represent the significance value of pixels. The pixels with the same color value c_j are grouped together to obtain the saliency of each color:

$$\begin{aligned} S(I_k) &= S(C_i) \\ &= \sum_{j=1}^{i} f_i D(c_i, c_j), \end{aligned} \tag{13}$$

where c_j represents the color value of pixel I_k, k represents the total number of colors contained in the image, and f_i represents the probability of c_j appearing in the image [25–27].

Histogram-based acceleration: In order to speed up the calculation, the number of colors to be calculated is reduced from 2563 to 123 by quantizing 256 to 12 color values per color channel. Then, the color histogram of the input image is calculated, and only the high-frequency colors are retained. The remaining colors that appear less frequently are replaced by the nearest color in the histogram [23, 28].

Color space smoothing: To reduce the error generated when color quantization occurs, the significance value of each color is replaced by a weighted average of the significance of similar colors. After experimental verification, it can produce better results when quantizing in RGB space and measuring distance in Lab space. The saliency detection algorithm based on histogram contrast is very simple and easy to understand in theory. It calculates saliency according to global contrast and has a fast calculation speed. It has a good effect on image resources with less complex background information [24].

3. Saliency Detection Module of the Algorithm in This Paper

3.1. Global Saliency Detection Module. The global saliency detection module mainly extracts the context information of the image. The input of this module is P2 with the size of $256 \times 256 \times 256$ generated by the feature extraction module and ITF, the image and text embedding vector generated from the image description branch. P2 has rich contextual information, and ITF has bimodal information and more attention to prominent areas. The two inputs ensure the validity and completeness of the module's information. For IFT, the number of channels is changed from 1024 to 256 by 1×1 convolution, and then the amount is changed to 3d tensor and the size is $256 \times 256 \times 256$ by stacking operation. Stacking operation is to copy $1 \times 1 \times 25\,256 \times 256$ times and cascade the generated 3D tensor with P2 through convolution and then generate a feature graph with the size of $256 \times 256 \times 256$ through two convolution layers. After a full connection of the feature graph, the result of global significance detection module is obtained. The above process can be described by the following formula:

$$G = \mathrm{ReLU}\left(W_p\left(\mathrm{Cat}\left(W_{p1}P2 + b_{p1}\right), \mathrm{Tile}\left(W_{p2}I + b_{p2}\right)\right) + b_p\right), \tag{14}$$

where W_p, W_{p1}, and W_{p2} are the weight parameters of the convolution layer, b_p, b_{p1}, and b_{p2} are the bias parameters of the convolution layer, I represents the embedded vector of text and text, P2 is the three-dimensional tensor of the shared feature network, Cat() represents the cascade operation, ReLU is the activation function, and G represents the generated feature with a size of $256 \times 256 \times 256$, as shown in the following formula:

$$\mathrm{GSM} = \mathrm{Sigmoid}\left(W_s G + b_s\right). \tag{15}$$

G contains information in RGB and text modes, and a saliency graph is generated through convolution operation and Sigmoid function. W_s and b_s represent the weight parameters and bias parameters of the convolution layer, respectively, Sigmoid is the activation function, and Global Saliency Mask (GSM) is the Global Saliency map with a size of $256 \times 256 \times 2$.

3.2. Local Significance Detection Module. In this paper, using the target detection technology to obtain local information, first of all, to locate objects in the image, generate the corresponding target box, and then to locate the target object segmentation in the frame, such that neither can destroy the salient objects within the structure, keep the consistency, also can improve the accuracy of local information, there will be no problem of the missing target.

In the local significance detection module, an RGB image is input, and a series of bounding boxes are generated using Region Proposal Network (RPN) structure. In order to effectively use multiscale information to assist target detection, the feature pyramid network structure is adopted at the head of CNN Network to integrate the $\{P_i\}_{i=2}^{5}$ information of the feature extraction module and generate the corresponding candidate box $\{B_i\}_{i=1}^{N}$, where N represents the preset number of candidate boxes for each image and the feature of candidate boxes is $\{f_{B,i}\}_{i=1}^{N}$.

There are two branches in this module. One branch outputs the probability of significant target segmentation and the other branch outputs the probability of candidate box target regression and classification, which is referred to

as the candidate box recognition branch. In the candidate box identification branch, feature $f_{B,i}$ is first convolved through a layer to output feature f_{B1} with the size of $7 \times 7 \times 256$, and then vector f_{B2} with the size of 1024 is obtained through a convolution layer. After that, the ITF embedded vector from the image description branch is fused. The fusion operation adopted in this module is cascaded. After being cascaded, the number of channels is changed through 1×1 convolution to speed up network training. The image embedding vector provides the category information of the image, which can assist the screening of salience candidate boxes, and the selected candidate boxes will be splinted into the final salient prediction graph.

The above calculation process is described as formula (16). W_B and b_B represent the weight parameters and bias parameters of the convolution layer, respectively, and I represents the graph-text embedding vector ITF of the image description. The generated feature L predicts the classification probability and regression probability of the candidate box through two fully connected layers. If the classification probability is greater than the preset threshold θ, it will be selected as the salience candidate box, and then the candidate box will be matched to the position of the input image through mapping.

$$L = \mathrm{ReLU}\left(W_B\left(\mathrm{Cat}\left(f_{B2}, I\right)\right) + b_B\right). \tag{16}$$

In saliency prediction branches, the generated features are directly passed through the two convolution layers to generate candidate box prediction graphs with size of $28 \times 28 \times 2$. Combined with the classification information of the other branch, candidate boxes that are significant targets are screened out to output the final saliency prediction graph.

3.3. Element Distribution. An area is considered special or unique if it is salient in the whole image and not otherwise. It is generally believed that in the whole image, most of the widely distributed areas belong to the features of the background, while the foreground area, that is, the prominent target, is more concentrated and compact in its spatial distribution. The more compact the image features are distributed in the whole image space, the higher the salience value of this region will be.

The distribution of elements is similar to the uniqueness of elements. According to the distribution of features, a distribution feature is defined to calculate the distribution of features in this region. In this paper, D_i is used to represent the spatial distribution of features of pixel i, and the formula is as follows:

$$D_i = \sum_{j=1}^{N} \left\| p_j - u_i \right\|^2 \cdot w\left(k_i - k_j\right). \tag{17}$$

The above formula u_i is the weighted average position of the features.

$$u_i = \sum_{j=1}^{N} w\left(k_i - k_j\right) p_j. \tag{18}$$

In order to more truly reflect the spatial distribution of features, it is necessary to consider the influence of similar features. Therefore, the influence of position should be considered in the process of feature calculation, and the weight $w(k_i - k_j)$ is introduced. The formula is defined as follows:

$$w\left(k_i - k_j\right) = \frac{1}{Z_i} \exp\left(-\frac{1}{2\sigma^2}\right) \left\| k_i - k_j \right\|^2. \tag{19}$$

D_i in formula (17) above is used to measure spatial distribution. By calculating the distance between pixel features of a point and its average position, element distribution calculates the spatial distribution of element features and can more accurately determine the significance of the region.

3.4. Spatial Contrast Characteristics. In the process of calculation, the need for proper pixel level will be saliency assigned to the fine details, namely USES; the nonlinear fusion method obtains more accurate figure; the uniqueness and distribution image characteristics, in the form of nonlinear function, are used to the uniqueness of elements and spatial distribution characteristics of fusion; in this way, better calculation can contrast feature space. The formula is as follows:

$$S_i = U_i \cdot \exp\left(-k \cdot D_i\right). \tag{20}$$

In formula (20), U_i is the uniqueness value of computed pixels, D_i is the distribution value of computed pixels, and the two are normalized to [0, 1]. k is set to 6 according to the empirical value.

3.5. Fusion Mechanism. Saliency detection generally generates a variety of feature maps based on different features, but each feature map has its own limitations and can better calculate saliency in a certain aspect while relying on a single feature detection effect is relatively poor. Therefore, the shortcomings of a single feature graph can be improved in other different feature graphs, and multiple feature graphs can be combined to form complementary advantages so as to better improve the quality of generated salient graphs. Cellular Automata (CA) is a simple dynamic system that can simulate different features of images. In this model, a cell represents a pixel, and the value of a cell is defined as the value of a pixel point. The next state of each cell is determined by the current state of itself and its neighbors, and the significant value of each cell represents the current state of the pixel.

In this paper, multilayer cellular automata (MCA) is used to fuse saliency detection features. MCA is an effective mechanism to fuse multiple saliency graphs, and the pixels with similar or the same position in different saliency graphs are called neighbors. This mechanism shows complex and effective global characteristics. In this method, each cell displays only one state at a given time, and this state interacts with the states of neighboring cells. This mechanism can traverse saliency graphs of different features to each layer of cellular automata and give full play to the advantages of

saliency graphs of each feature so as to get saliency graphs with better fusion effect.

In every cell fusion mechanism, said a pixel of the image, the significant value by cellular automata in a different state of a significant figure in the set of pixels with the same coordinates with each other is called neighbors, namely, in any significant figure in a pixel, it has $M-1$ on other significant figure neighbors and defines all the neighbors when determining the next state of cellular automata has the same influence, In this paper, the method of MCA mechanism is adopted to obtain the fused saliency map by adopting the advantages of frequency domain feature, spatial contrast feature, and color feature, respectively.

The final significance graph formula is as follows:

$$S = \frac{1}{M}\sum_{j=1}^{N} S_m. \tag{21}$$

In formula (21), M represents the number of fused salient images. In this paper, the frequency feature, color feature, and spatial contrast feature of images are adopted, and MCA mechanism is adopted for feature fusion. Therefore, $M = 3$, $S_m = [S_{\text{frequencydomain}}, S_{\text{color}}, S_{\text{spatialcontrast}}]$ in this paper are frequency feature map, color feature map, and spatial contrast feature map, respectively.

4. Experiment and Analysis

4.1. Data Set and Evaluation Criteria. The experiment was conducted through three data sets, MSRA, SED2, and ECSSD. Among them, these three data sets all contain the original image and the corresponding Ground Truth (GT) graph, which is the real significance area of the image generated by artificial pixel recognition.

The MSRA dataset contains 1000 images, which are relatively simple structures, consisting of a simple background and a prominent object in the middle of the image, and the salient object to be segmented is a single object. This dataset is widely used, and almost all significance algorithms have been evaluated by using the MSRA dataset.

SED2 data set contains 100 images, the data set of images and MSRA data set, the image also contains a salient object, but the difference between the two is that the significant SED2 data set object is no longer just in the middle part of the image, salient objects of variable size and random in a different location in the image.

The ECSSD dataset is much more complex than the MSRA and SED2 datasets. This dataset contains 1000 images, all of which are completely random, including the number, position, and size of significant objects in the image are random, and the image background is extremely complex, which makes the recognition process of the recognition algorithm extremely challenging.

In this experiment, the Precision-Recall curve (PR curve) was adopted for the significant images recognized by the algorithm. Among them, the accuracy rate refers to the ratio of all the salient images identified by the algorithm and the individuals correctly predicted the correct salient images; the recall rate refers to the ratio of all the salient images identified by the algorithm and the individuals correctly predicted to the salient images that are actually correct.

The experiment also draws the Receiver Operating Characteristic curve (ROC curve), where the abscissa of the curve is a false positive rate (FPR), which means that among all the samples that are actually negative, they are wrongly judged as Positive. The calculation formula is as follows:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{22}$$

The true positive rate (TPR) of this curve represents the proportion of all samples that are actually Positive and are correctly judged to be Positive, which can be calculated as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{23}$$

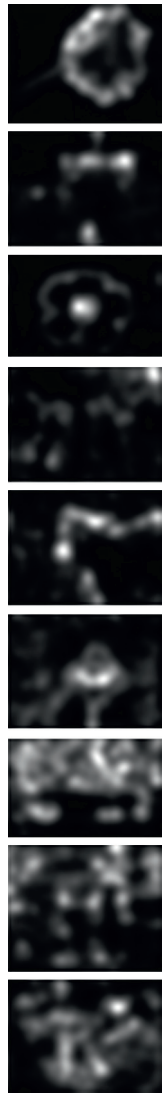
The effect of the algorithm is further analyzed through the curve.

4.2. Experimental Data and Analysis. The proposed algorithm and COV, FT, HC, Itti, LC, and SR visual saliency detection algorithms were tested on MSRA, SED2, and ECSSD datasets using the above evaluation methods. Three images are selected from each dataset, and the segmentation results of each algorithm are shown in Figure 1.

In Figure 1, rows 1 to 3 are MSRA databases, rows 4 to 6 are SED2 databases, and rows 7 to 9 are ECSSD databases. (a) is the original database map, (b) to (h) are algorithm recognition images, respectively: Ours, COV, FT, HC, Itti, LC, SR. A preliminary judgment can be made only through subjective analysis of the saliency image at the operation of the algorithm. Compared with the original image, the image run by the algorithm in this paper can well identify the saliency region, the saliency object is complete, and the specific image can be seen. In contrast, the image recognized by COV is very fuzzy. Only the salient region in the image is recognized, the salient object cannot be distinguished, and the contour is also very fuzzy. FT performs better and can recognize most of the images; some can even clearly identify what the salient object is, but there are some individual images and backgrounds mixed together; it is difficult to identify specific things. HC also has a good operation effect, similar to FT, and some images can be clearly identified, but there are also some general image effects, relatively fuzzy and very similar to the background, difficult to identify. The operation effect of Itti is general. Only a few images can distinguish what the object is, and most of them are very fuzzy and cannot be distinguished, but the saliency area is well presented. LC performs well in some image recognition, but some performance is mediocre. In addition, the image recognized by LC is mixed with some background information, so the recognition of salient region is not specific enough. Finally, SR has a poor recognition effect. It fails to clearly identify the salient area of the image, and some images are even difficult to identify in the dark. Some of the recognized images are also very fuzzy. The performance of all algorithms on three databases is compared and analyzed

(a) (b) (c) (d) (e) (f) (g)

Figure 1: Continued.



(h)

Figure 1: Comparison results of image segmentation.

from the visual saliency images, and the effect of the algorithm is generally understood. Then more accurate analysis will be carried out through specific image curves. PR images drawn by experiments on all three data sets are shown in Figures 2–4.

In terms of the overall trend of the image, except for the performance of FT and HC on the ECSSD data set, the performance of all other algorithms on the three data sets shows that: (1) when the recall rate is greater than a certain value, the precision rate of the algorithms on each data set gradually tends to be flat with a small fluctuation range. (2) When the recall rate is greater than a certain value, the precision rate of all algorithms begins to decline.

As shown in Figure 2, on the MSRA data set, it can be seen that the algorithm in this paper has the best effect. With the same recall rate, the accuracy of the algorithm in this paper is higher than all other algorithms, and when the recall rate reaches 0.5, the precision rate reaches the highest value, which can reach 0.57. Secondly, the better effect is COV and SR. Compared with the two algorithms, SR has a higher accuracy when the recall rate is low, up to nearly 0.4, but with the increase in recall rate, the accuracy keeps decreasing. On the contrary, when the recall rate is 0.1, the accuracy of the COV is very low, but with the increase of the recall rate, the accuracy keeps rising, and when the recall rate reaches the range of 0.5-0.6, the accuracy reaches the peak of about 0.44. Among them, Itti and HC have poor performance. The peak accuracy of Itti is only about 0.25, and HC also has poor overall performance. However, when the recall rate is about 0.17, the peak accuracy can barely reach about 0.28.

As shown in Figure 3, on the SED2 data set, the performance of each algorithm is significantly improved comprehensively compared with that of the MSRA data set. The highest accuracy of the algorithm is about 0.65, which is 0.1 higher than the MSRA dataset of 0.55.

In terms of the performance of specific algorithms, the algorithm in this paper has the best effect, and FT also has a good effect. When the recall rate of the algorithm in this

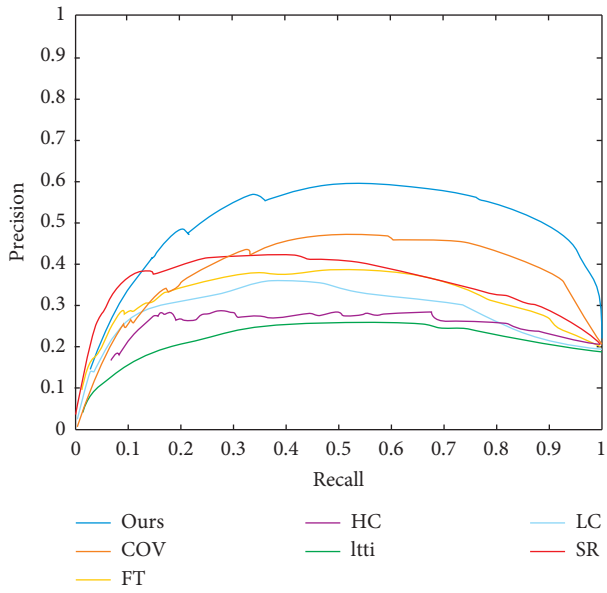


Figure 2: PR diagram on MSRA dataset.

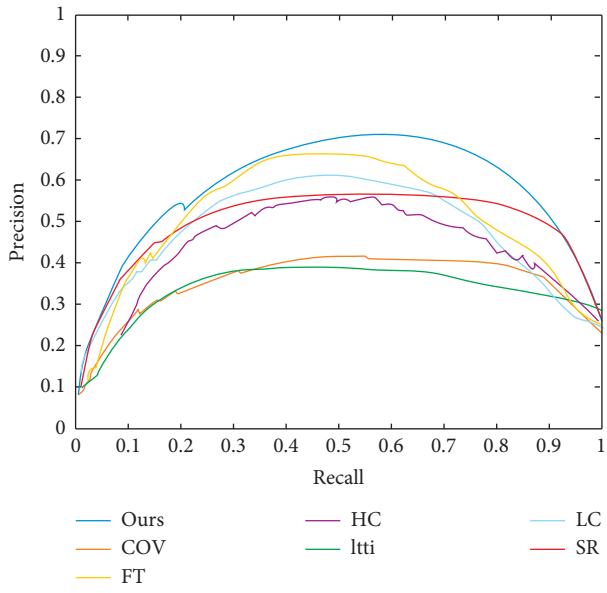


Figure 3: PR diagram on SED2 dataset.

paper is 0.6, the accuracy rate reaches the highest 0.65. FT also achieves 0.65 at the recall rate of 0.4, which is the same as the highest accuracy of the two algorithms. However, the algorithm in this paper has a larger overall duration interval, which is within the range of recall rate of 0.33 to 0.8. The accuracy of the algorithm in this paper is above 0.6, while the recall rate range of FT is only from 0.35 to 0.65.

On the SED2 data set, some algorithms also have mediocre performance. Among them, Itti has the worst effect, and its accuracy is far different from other algorithms. The highest accuracy of Itti is only about 0.32, which is 0.14 worse than that of COV, which is the second to last. Although the effect of COV is much better than that of the worst Itti, it is still much different from other algorithms on the whole. In addition, on the SED2 data set, all algorithms reach the peak accuracy when the recall rate reaches 0.5 to 0.6.

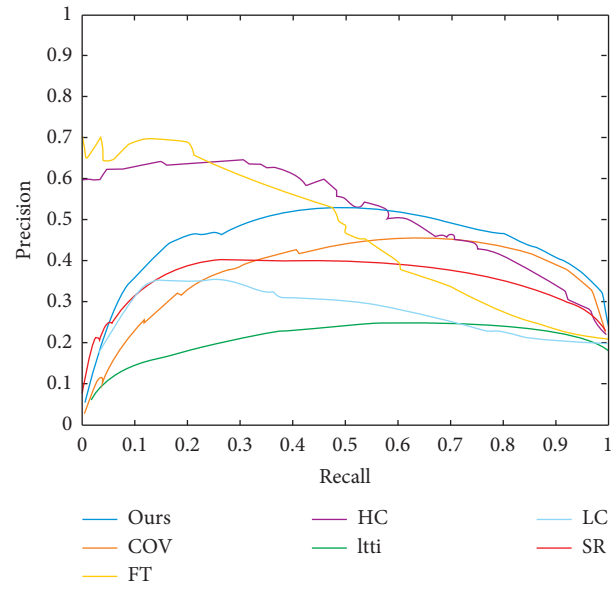


Figure 4: PR graph of all algorithms on ECSSD dataset.

As shown in Figure 4, the trend of FT and HC on the ECSSD data set is different from all other algorithms, including the trend of each algorithm on the other two data sets. When the recall rate is zero, the accuracy of FT is as high as 0.7, which is also the highest accuracy of all algorithms in the ECSSD data set. When the recall rate is zero, HC also achieves an accuracy of nearly 0.6, which is much higher than other algorithms. However, the accuracy of FT and HC decreases with the decrease in recall rate. When the recall rate is greater than 0.5, the accuracy of FT starts to be lower than that of the proposed algorithm and COV; when the recall rate of HC is greater than 0.6, the accuracy of HC starts to be lower than that of the proposed algorithm and COV.

In this data set ECSSD, Itti still performs the worst. The highest accuracy of Itti on this data set is only about 0.25, far lower than that of other algorithms. However, the curve of Itti is relatively smooth, and the accuracy is balanced with the fluctuation of the recall rate. According to the PR graph, the maximum accuracy of each algorithm on three data sets was obtained, as shown in Table 1.

When it comes to the MSRA data set, the algorithm in this paper has the highest accuracy of 0.56, followed by SR with similar accuracy of 0.40 and COV with 0.43. The two algorithms with the worst performance were Itti and HC, with an accuracy of 0.26.

On the SED2 data set, the two algorithms with the best performance are the algorithm in this paper and the FT. The accuracy of the two algorithms is similar, the accuracy of the former is 0.64, and the accuracy of the latter is 0.62, with a 0.02 difference. The least effective algorithm was Itti, with an accuracy of 0.35, but the highest accuracy of Itti on the SED2 dataset was slightly higher than its value on the MSRA dataset. In addition, the highest quasi exclusion rate of HC with general performance on the MSRA data set increased

Table 1: The highest accuracy of each algorithm and average accuracy of each algorithm on each data set.

Algorithm	MSRA	SED2	ECSSD	Average accuracy
Itti	0.26	0.35	0.26	0.290
SR	0.40	0.57	0.37	0.447
FT	0.38	0.62	0.69	0.563
LC	0.37	0.59	0.39	0.450
HC	0.28	0.58	0.60	0.487
COV	0.43	0.46	0.49	0.460
Ours	0.56	0.64	0.53	0.577

significantly, from 0.28 on the MSRA data set to 0.58 on the SED2 data set. The maximum accuracy of each algorithm on the SED2 data set has all increased, and the increase is large.

On the ECSSD data set, the accuracy of the ECSSD algorithm is 0.69, nearly 0.7, which is the highest value of all algorithms on the three data sets. In addition, compared with the performance of all algorithms on the previous two data sets, it is found that the performance of FT, HC, and COV on this data set is better. The accuracy of Itti is not ideal.

By integrating the highest accuracy of each algorithm in the three data sets, it can be seen that Itti, COV, and CA have little fluctuation. That is, the performance of the algorithm and the complexity of image resources have little influence. By averaging the highest accuracy of each algorithm on three data sets, the average highest accuracy of each data set is obtained. It can be seen that the algorithm in this paper has the highest accuracy, and the algorithm in this paper has the best image recognition effect, followed by FT. Itti has the lowest accuracy in image recognition.

Each algorithm was further analyzed by ROC curve, as shown in Figures 5–7. In order to quantify the ROC curve presented by the algorithm for comparison, a new variable, the area under ROC curve (AUC), is introduced here. This variable is the amount of area under the curve. Generally, the ROC curve is located above the straight line $y = x$. Therefore, AUC values are between 0.5 and 1. The larger the AUC value is, the better the effect of the model is.

As shown in Figure 5, on the MSRA data set, the algorithm in this paper and COV perform well, and the AUC values of the two algorithms are significantly higher than those of other algorithms, which indicates that when the error rate of the abscissa algorithm is constant, the two algorithms have higher accuracy. In addition, HC performed poorly on the dataset, with the minimum AUC value and HC algorithm curve tending to $y = x$ line, indicating that the algorithm segmented the image and the number of significant images guessed to be correct and wrong was split 50/50.

As shown in Figure 6, on the SED2 data set, it can be seen that the AUC value of the algorithm in this paper is obviously the largest, and the curve trend of other algorithms is below the curve of this algorithm, so the algorithm has the best effect on this data set. The curve of Itti is lower than that of other algorithms. The AUC value of Itti is the smallest, and the effect of Itti is relatively poor.

As shown in Figure 7, on the ECSSD data set, the algorithm is not very concentrated, and the effect difference is

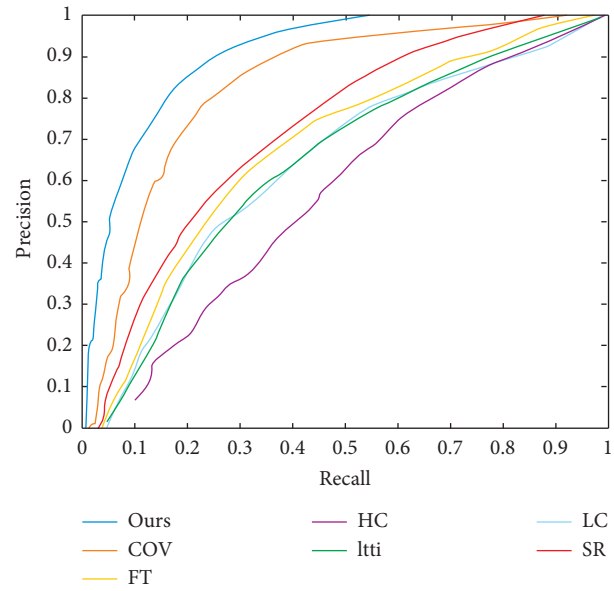


Figure 5: ROC diagram of all algorithms on MSRA dataset.

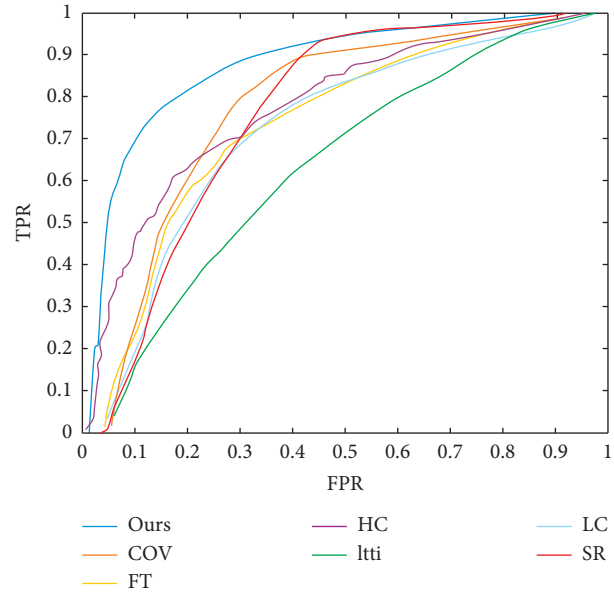


Figure 6: ROC diagram of all algorithms on SED2 dataset.

larger than that of the first two data sets. In this data set, the results of the proposed algorithm and COV are good, the curve is above other algorithms, and the AUC area is the largest. FT is the second most effective.

According to the above research, it can be seen that among the saliency images generated by various algorithms, the effects are not the same. Some algorithms can well identify the saliency region, but the recognized object is difficult to identify, such as COV. Some algorithms can well identify the saliency region and saliency object, but some background information will be mixed, such as Itti and HC.

Itti and HC, the mixed algorithm of background information, can better convey the salient regional characteristics and convey information more completely; the

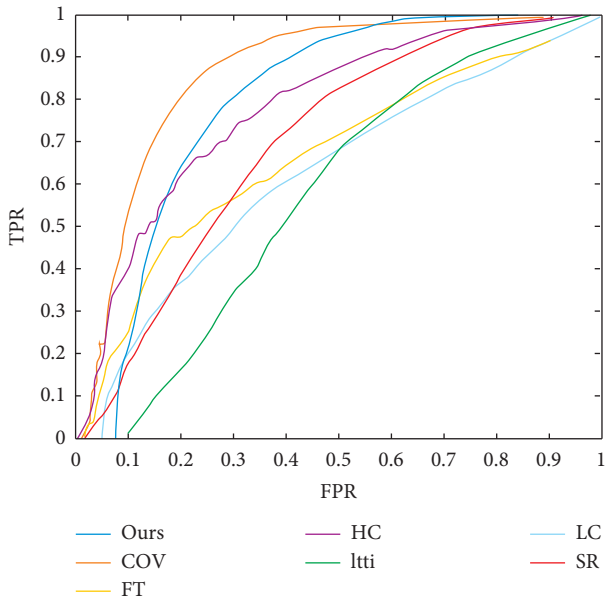


Figure 7: ROC diagram of all algorithms on the ECSSD dataset.

algorithm can be applied in the related areas of image recognition, by significant objects, and the combination of the relevant background information makes convey more comprehensive, letting the computer image identification be more accurate. Similar to the COV, the salient area of the object in the image is very accurate, but the specific contour of the object is difficult to identify. Similar to the algorithm, it can be used in the field of computer vision positioning. Although the contour of the specific object cannot be seen but used for positioning, it is enough to show the salient area of the image.

In addition to the obvious image effect generated by the algorithm, it is also necessary to consider the running speed of the algorithm itself. If the algorithm is very accurate even if it is recognized in some fields requiring fast image processing by the computer, if it runs slowly, the effect will be greatly discounted.

FT has good accuracy. That is, it can well identify the salient region in the image. However, due to the obstacles of its running speed, this algorithm cannot be applied to the field of image recognition and image segmentation. However, the algorithm in this paper has a high accuracy of image recognition and a high speed of algorithm operation, so the algorithm proposed in this paper can be well applied to the field of image segmentation and image recognition.

5. Conclusion and Future Work

Image saliency detection in the computer vision field has a great role in promoting the development of this paper for the image saliency detection algorithm based on the attention mechanism studied and then introduced the visual attention mechanism among some of the basic principles and common salient features; it also introduced some of the commonly used detection model, publicly available data sets, and evaluation index, etc. Based on the above research, this paper proposes an image object detection method that integrates global and local information. Firstly, the edge operator is used to remove the nodes near the boundary to obtain the preprocessing graph. Then, the local saliency map was calculated by fusing a multiscale superpixel segmentation saliency map, and the initial saliency map was obtained by fusing the global saliency map and local saliency map based on the Hadamard product. Furthermore, the initial saliency map was optimized by conditional random fields to obtain a saliency map with smoother boundary information. Then the center prior algorithm and edge prior algorithm are, respectively, used to obtain the center prior map and edge prior map. Finally, the fusion optimization process fuses three feature maps and two prior maps to generate the final saliency map. Experimental results show that the proposed algorithm can effectively suppress the interference of complex background information and solve the problem of prominent object contour blur.

Future research work is as follows:

(1) When extracting contrast features, the proposed algorithm uses a convex hull to roughly locate the image, which may introduce more background areas in the image, thus reducing the detection effect of contrast feature maps.

(2) When extracting a texture feature map, the algorithm in this paper can obtain a texture feature map with finer texture by gradually tuning weight. Therefore, in the future work, we will further consider how to accurately locate the image so as to obtain the final saliency map with better detection results.

(3) The algorithm in this paper adopts the Hadamard product matrix to fuse the feature information, which cannot fully fuse the local edge information and global salience information. On the other hand, the computational complexity is large when searching for the optimal threshold value. Therefore, it can be considered how to construct a fusion model integrating global information and local information by using the Bayesian criterion. Make full use of feature information to further refine the boundary contour of prominent objects.

Data Availability

The authors confirm that the data supporting the findings of this study are available within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key Research and Development Program of China during the 13the Five-Year Plan Period (No. 2016YFD0701003).

References

[1] R. M. Cong, J. J. Lei, and Q. M. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2941–2959, 2019.

[2] Y. Ji, H. Zhang, and Q. M. Wu, "Saliency detection via conditional adversarial image-to-image network," *Neurocomputing*, vol. 316, pp. 357–368, 2018.

[3] L. Ye, Z. Liu, X. Zhou, L. Shen, and J. Zhang, "Saliency detection via similar image retrieval," *IEEE Signal Processing Letters*, vol. 23, no. 6, pp. 838–842, 2016.

[4] Y. Z. Niu, L. N. Lin, Y. Z. Chen, and L. L. Ke, "Machine learning-based framework for saliency detection in distorted images," *Multimedia Tools and Applications*, vol. 76, no. 24, pp. 26329–26353, 2017.

[5] G. Lin, J. F. Zhao, and Y. T. Chen, "Fast full resolution saliency detection based on incoherent imaging system," *Optical Review*, vol. 23, no. 4, pp. 601–613, 2016.

[6] R. M. Cong, J. J. Lei, H. Z. Fu, Q. M. Huang, and X. C. Cao, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 568–579, 2018.

[7] J. R. Ren, Z. Liu, X. F. Zhou, G. L. Sun, and C. Bai, "Saliency integration driven by similar images," *Journal of Visual Communication and Image Representation*, vol. 50, pp. 227–236, 2018.

[8] J. F. Guo, T. W. Ren, L. Huang, and J. Bei, "Saliency detection on sampled images for tag ranking," *Multimedia Systems*, vol. 25, no. 1, pp. 35–47, 2019.

[9] N. Rabbani, B. Nazari, S. Sadri, and R. Rikhtehgaran, "Efficient Bayesian approach to saliency detection based on Dirichlet process mixture," *IET Image Processing*, vol. 11, no. 11, pp. 1103–1113, 2017.

[10] Z. G. Jin, J. K. Li, and D. Li, "Co-saliency detection for RGBD images based on effective propagation mechanism," *IEEE Access*, vol. 7, pp. 141311–141318, 2019.

[11] Z. W. B. Liu, L. N. Zou, L. Li, S. Liquan, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Processing Letters*, vol. 21, pp. 88–92, 2014.

[12] C. Ma, G. H. Gu, M. J. Wan, C. C. Song, and H. Zhang, "Infrared small target detection based on fusion of multiple saliency information," *Applications of Digital Image Processing Xliii*, vol. 11510, 2020.

[13] Z. Li, C. Y. Lang, J. S. Feng, Y. D. Li, T. Wang, and S. H. Feng, "Co-saliency detection with graph matching," *Acm Transactions on Intelligent Systems and Technology*, vol. 10, no. 3, 2019.

[14] X. F. Zhang, Y. Wang, and D. H. Wang, "Saliency detection via image sparse representation and color features combination," *Multimedia Tools and Applications*, vol. 79, no. 31-32, pp. 23147–23159, 2020.

[15] S. Bardhan, S. Das, and S. Jacob, "Visual saliency detection via convolutional gated recurrent units," in *Proceedings of the 26th International Conference on Neural Information Processing (ICONIP) of the Asia-Pacific-Neural-Network-Society (APNNS), Neural Information Processing (ICONIP 2019)*, pp. 162–174, Syndey, Australia, December 2019.

[16] R. M. Cong, J. J. Lei, H. Z. Fu, J. H. Hou, Q. M. Huang, and S. Kwong, "Going from RGB to rgbd saliency: a depth-guided transformation model," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3627–3639, 2020.

[17] Z. H. Chen, Y. Liu, B. Sheng, J. N. Liang, J. Zhang, and Y. B. Yuan, "Image saliency detection using gabor texture cues," *Multimedia Tools and Applications*, vol. 75, no. 24, pp. 16943–16958, 2016.

[18] C. P. Li, Z. X. Chen, Q. M. J. Wu, and C. Y. Liu, "Saliency object detection: integrating reconstruction and prior," *Machine Vision and Applications*, vol. 30, no. 3, pp. 397–406, 2019.

[19] Z. Z. Tu, T. Xia, and J. Tang, "RGB-T image saliency detection via collaborative graph learning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 160–173, 2020.

[20] H. K. Song, Z. Liu, Y. F. Xie, L. S. Wu, and M. K. Huang, "RGBD Co-saliency detection via bagging-based clustering," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1722–1726, 2016.

[21] R. Huang, W. Feng, and Y. B. Zou, "Exemplar-based image saliency and Co-saliency detection," *Neurocomputing*, vol. 371, pp. 147–157, 2020.

[22] L. Li, F. G. Zhou, Y. Zheng, and X. Z. Bai, "Reconstructed saliency for infrared pedestrian images," *IEEE Access*, vol. 7, pp. 42652–42663, 2019.

[23] X. F. Zhang, Y. Wang, J. Yan, Z. X. Chen, and D. H. Wang, "A unified saliency detection framework for visible and infrared images," *Multimedia Tools and Applications*, vol. 79, no. 25-26, pp. 17331–17348, 2018.

[24] H. J. Li, C. B. Li, and Y. P. Ding, "Fall detection based on fused saliency maps," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 1883–1900, 2021.

[25] H. K. Yu, K. Zheng, J. W. Fang, H. Guo, and S. Wang, "A new method and benchmark for detecting Co-saliency within a single image," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3051–3063, 2020.

[26] Q. M. Peng, Y. M. Cheung, X. G. You, and Y. Y. Tang, "A hybrid of local and global saliencies for detecting image salient region and appearance," *IEEE Transactions on Systems Man Cybernetics-Systems*, vol. 47, no. 1, pp. 86–97, 2017.

[27] Y. Zheng, F. G. Zhou, and C. M. Sun, "Mutual guidance-based saliency propagation for infrared pedestrian images," *IEEE Access*, vol. 7, pp. 113355–113371, 2019.

[28] C. B. Zhu, W. H. Zhang, T. H. Li, S. Li, and G. Li, "Exploiting the value of the center-dark channel prior for salient object detection," *Acm Transactions on Intelligent Systems and Technology*, vol. 10, no. 3, 2019.