This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3179461, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

Visual Saliency Detection in High-Resolution Remote Sensing Images using Object-Oriented Random Walk Model

Lin Ding, Xing Wang, and Deren Li

***Abstract*—As high-resolution remote sensing images begin to integrate new characteristics such as a great volume of data, a wide variety of ground objects and a high structural complexity, traditional methods previously used for feature extraction in low-resolution remote sensing images are inefficient and inadequate for the accurate feature description of various objects. Thus, object feature extraction from high-resolution remote sensing image remains to be a challenge. To address this issue, we introduced the visual attention mechanism into high-resolution remote sensing image analysis in this study by proposing a novel object-oriented random walk model for visual saliency (ORWVS) detection from high-resolution remote sensing images. In the proposed model, an object-oriented random walk strategy is designed to simulate the transfer path of visual focus on the images, and to extract the local salient regions in an efficient and accurate manner, laying a foundation for accurate feature descriptors. The ORWVS model is compared to eight visual attention models, and the experiments prove its superiority.**

***Index Terms*—focus of attention (FOA), salient object detection, random walk, visual saliency**

I. Introduction

With the continuing development of new sensor technology and earth observation technologies, spatial resolution of remote sensing images continues to increase. At present, the spatial resolution of GeoEye, an international commercial remote sensing satellite, has reached 0.41 meters, the U.S. military reconnaissance satellite KH-12 has reached an optical resolution of 0.1 meters [1], and the spatial resolution of the Chinese Earth observation satellite Gaofen-2 has reached 0.8 meters [2]. These high-resolution remote sensing images document detailed and complex land covers with rich color (spectrum), texture, geometric and structural information, offering us a new opportunity for advancing the interpretation of remote sensing images. At the same time, geospatial object detection and scene-level geographic image categorization, the two fundamental yet challenging research aspects of remote sensing image analysis, have attracted increasing attention.

Traditional pixel-based analysis methods for low-resolution remote sensing images are not powerful enough when dealing with high-resolution remote sensing images. They also fall short in obtaining accurate and discriminative descriptions of objects.

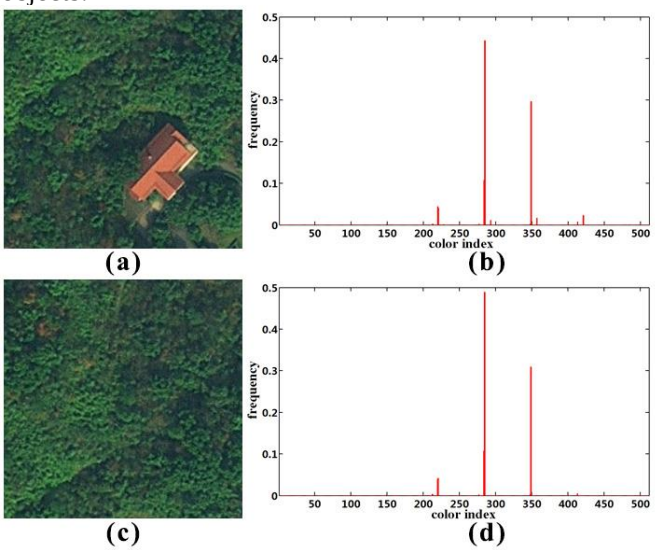


Fig. 1. Comparison of global color histogram of remote sensing images. (a) WorldView-2 remote sensing image 1. (b) Color histogram (quantized to 512 levels) of image 1 in the CIELab color space. (c) WorldView-2 remote sensing image 2. (d) Color histogram (quantized to 512 levels) of image 2 in the CIELab color space.

Fig. 1 shows a set of examples where (a) and (c) present two WorldView-2 remote sensing images and (b) and (d) demonstrate their corresponding global color histograms (quantized to 512 levels) in the CIELab color space. The major difference between images (a) and (c) is the existence of a house with an orange roof in (a). Despite the fact that the local salient objects in the two images own great visual differences, the color histogram distributions of the two images are very similar to each other due to the small proportion of the changing area (i.e., the house) in (a). Therefore, for the remote sensing image (a), traditional feature extraction methods that fail to separate the image target from the background are unsuitable to represent local salient objects. However, when facing high-resolution

This work was supported in part by the National Natural Science Foundation of China under Grants 42090012, in part by 03 special research and 5G project of Jiangxi Province in China (20212ABC03A09); Zhuhai industry university research cooperation project of China (ZH22017001210098PWC); Sichuan Science and Technology Program (2022YFN0031) and Zhizhuo Research Fund on Spatial-Temporal Artificial Intelligence (Grant No.ZZJJ202202). *(Corresponding author: Lin Ding)*.

Lin Ding and Deren Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: dinglin@whu.edu.cn; drli@whu.edu.cn).

Xing Wang is with the School of Marine Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: xing.wang@tju.edu.cn).

remote sensing images, humans are often able to discern the characteristics of these ground objects in a rapid and accurate manner. Thus, how to simulate the human visual attention mechanism to obtain visual features of various ground objects remains to be challenging task.

As early as the 1980s, neuroscientists discovered that when facing massive amounts of visual information in a complex scene, the human visual system (HVS) would selectively focus on some local regions and ignore the background that often takes a higher proportion area of the scene. By simulating the visual attention mechanism of HVS, scholars from the computer vision field came up with several visual attention models that include ITTI [3], GBVS [4], etc. The performances of these models have been validated in a number of natural image databases [5].

In this paper, the visual attention mechanism of HVS is introduced into high-resolution remote sensing image processing, and an object-oriented random walk model for visual saliency detection in remote sensing images is proposed. In the proposed model, an object-oriented random walk strategy is adopted to simulate the transfer path of visual focus on remote sensing images, aiming to obtain the local salient regions of images efficiently and accurately. This study lays a foundation for the accurate description of salient objects in high-resolution remote sensing images.

The remaining paper is organized as follows. Section II reviews the related works on visual attention models. Section III presents each stage of our visual saliency detection model. Section IV first introduces the datasets used for performance evaluation, and then shows the experimental results. Section V draws the conclusions.

II. RELATED WORK

Most existing efforts on the visual attention model (VAM) are concentrated in the field of natural image analysis, such as the ITTI model [3], a classical VAM proposed by Itti et al. from the University of Southern California in 1998 based on the classical "feature integration theory" [6]. The ITTI model contains five steps for calculating the scan path of focus of attention (FOA) in the image. First, an input image is subsampled into a Gaussian pyramid, with each pyramid level decomposed into channels for red, green, blue, yellow, intensity, and local orientations. Second, center-surround feature maps for different features are constructed and normalized from these channels. Third, maps are summed across the scale and normalized again in each channel. These maps are linearly summed and normalized once more to produce the "conspicuity maps". Fourth, the conspicuity maps are linearly combined to generate the saliency map. Finally, based on the saliency map, a "winner-take-all" (WTA) neural network [7], [8] and an "inhibition of return" method [9] are employed to obtain the scan path of FOA.

Inspired by the ITTI model, more VAMs were designed. Bruce and Tsotsos proposed the AIM (Attention based on Information Maximization) model by introducing the self-information metric in the classical Shannon information theory to image saliency calculation [10]. Harel et al. proposed a graph-based visual saliency (GBVS) computation method based on the ITTI model [4]. Hou and Zhang proposed a dynamic visual attention (DVA) model based on sparse features [11]. Garcia-Diaz et al. used local energy variability to measure the saliency of images with an adaptive whitening saliency (AWS) model [12]. Goferman et al. proposed a context-aware (CA) saliency detection model based on context awareness [13]. As the effectiveness of VAM in the field of image analysis has been confirmed through these studies, more attention has been paid to VAM with the design of other notable methods [5], [14].

From the perspective of human visual behavior, the selective attention mechanism of HVS often appears as the fixation and transfer of FOA. FOA is usually defined as the point that has the highest score of saliency in a scene. Therefore, the scan path of FOA is very important for the visual saliency distribution in the image. Studies in cognitive psychology have shown that the FOA transfer path has certain randomness, and an appropriate random walk model allows us to effectively predict the FOA transfer path [15], [16]. On the basis of this theory, scholars have proposed a series of VAMs based on random walk models for predicting the visual saliency distribution of images [4], [17], [18]. Among them, the most classical and influential model is the visual saliency model based on the graph theory proposed by Harel et al., called GBVS [4]. This model improves two steps on the basis of the existing models and traditional methods: one is the activation map generation, and the other is the activation map normalization and fusion. The GBVS model defines Markov chains in different feature maps, calculates the transition probability of the visual focus between the pixels by comparing the character difference and distance of pixel points, and treats the equilibrium distribution of FOA on the image pixels as the saliency value of the pixels. Experiments show that the GBVS model is able to accurately predict the fixation of FOA and owns notable advantages compared to other VAMs [19]. However, the GBVS model has some limitations. First, GBVS treats pixels as the basic unit and calculates the saliency value of each pixel by a constructed Markov chain. Such a Markov chain tends to have a large number of nodes, leading to great computational needs. Second, the GBVS model calculates the final saliency map with a Gaussian smoothing step, resulting in blurred edges of the salient region.

In light of the deficiencies of the GBVS model, this paper introduces the idea of object-based image analysis into the VAM field. The object consisting of adjacent similar pixels is used as the basic unit in visual saliency computing, as an alternative to a single pixel in the traditional way. An Object-oriented Random Walk model for Visual Saliency (ORWVS) detection in high-resolution remote sensing images is proposed. The ORWVS model not only reduces the number of nodes in the Markov chain but also obtains sharp edges from the saliency map, benefiting the extraction of salient regions from the images.

III. METHODOLOGY

Our proposed ORWVS model is mainly different from the existing GBVS model in that we introduce object-based image analysis for visual saliency map computation and regard objects

as the basic processing unit, which has been proved to be effective for high-resolution remote sensing image. Furthermore, we exploit random walk model to predict the transfer path of FOA between the objects in the image, and thus propose an object-based random walk strategy to calculate the visual saliency distribution of the objects. Fig.2 illustrates the process of our proposed ORWVS model, which include five steps: multi-scale segmentation, object feature extraction, FOA transition probability computation, visual saliency computation, and saliency maps fusion.

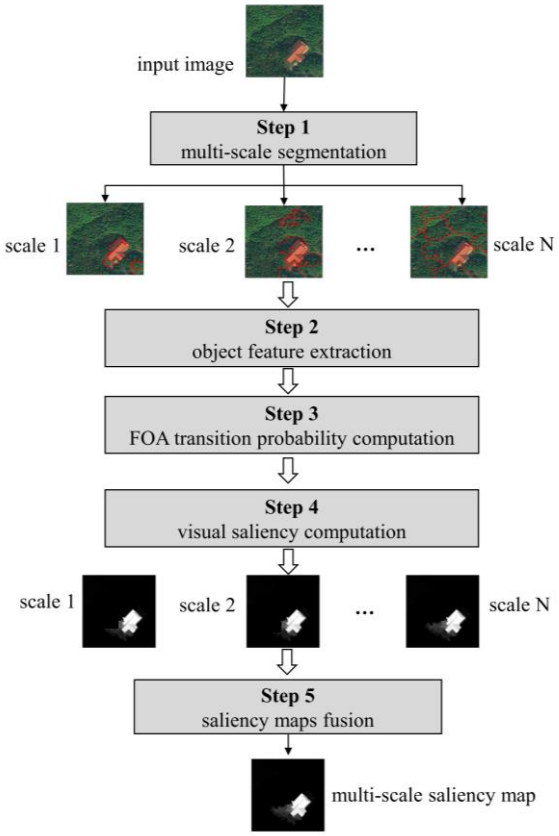


Fig. 2. The flowchart of the proposed ORWVS model

Specially, in the first step, N scales are set to perform multi-scale segmentation for the input image to obtain N segmentation maps, and the adjacent regions with similar color features in each segmentation map will be merged to improve segmentation result. Visual features including color, intensity and texture of each segmented region are then extracted to establish an object set for segmentation map of each scale. For the object set of each scale, the edge weights are derived from feature differences between the objects to calculate the transition probabilities of FOA, and the visual saliency map of each scale is then achieved via saliency computation, which are further fused to obtain the final saliency map.

A. Multi-scale Segmentation

The continuous improvement of spatial resolution of remote sensing images over the last decade facilitates the acquisition of detailed information about diverse ground objects. Such a trend redirected the research attention from pixel-based image analysis to object-oriented image analysis [1].

The idea of object-oriented image analysis was first applied in remote sensing image processing in the 1970s [21]. Since 2000, with the popularity of high-resolution remote sensing images, object-oriented image analysis methods have experienced rapid development due to their advantages [22-24]. The idea of object-oriented image analysis is to treat the object as the minimum image processing unit instead of pixels. A common approach for extracting objects from an image is image segmentation [20].

Image segmentation, a process of dividing an image into a number of homogeneous regions (also called "superpixels") without overlap [25], is an important task in the computer vision community. After years of development, great achievements have been made in image segmentation. The current mainstream image segmentation methods can be broadly divided into two categories, i.e., graph-based methods and gradient ascent methods [26].

To explore the performances of different image segmentation methods on visual saliency computing, three image segmentation methods were employed for comparative analysis, including the graph-based image segmentation (GS) [27], the quick shift (QS) method [28], and the simple linear iterative clustering (SLIC) method [26]. However, due to the wide variety of ground objects and the high structural complexity in high-resolution remote sensing images, different ground objects in images always appear on different scales. Therefore, it is challenging to extract various ground objects on a single scale. To address this issue, scholars have proposed solutions from two perspectives. Some solutions used multiple-scale parameters to perform multi-scale segmentation on remote sensing images; others performed an over-segmentation on remote sensing images, followed by merging over-segmented regions with certain constraints. Inspired by these solutions, we multiple-scale parameters to perform multi-scale segmentation on remote sensing images and further analyze the adjacency matrix and color features of the segmented regions at each scale, followed by a merging mechanism that merges the adjacent regions with similar color features.

It is worth noting that we perform multi-scale segmentation using N different scale levels where each level has its own scale, and the multi-scale parameter "N" is determined by manually setting one scale per level. Therefore, for each image, we can obtain N segmentation maps which are then further used for subsequent feature extraction for objects.

B. Object Feature Extraction

Image segmentation is intended to obtain a series of homogenous regions without overlap. However, if we want to construct objects on the basis of the segmented regions, the extraction of visual features from these segmented regions is needed. In the field of VAM, scholars often consider color, intensity, texture and orientation as the common visual features, with color features being the most widely used ones. In this study, we extract three types of common features from the objects, including color, intensity and texture, and provide a comparative analysis of these three types of visual features on the visual saliency computing performance in the following

experiment section.

(1) Color Feature Extraction

In early studies of VAMs, scholars generally used the red-green-blue (RGB) color space to obtain color feature maps. However, with the development in the color theory field, it was discovered that the RGB color space is inconsistent with the human perception of color psychology, as visual differences between two colors often failed to be presented accurately in the form of distance between two points in the RGB color space. To address this issue, many VAMs adopt visual perception-oriented color models, such as the hue-saturation-value (HSV) color space and the CIELab color space.

The HSV color space is a classic visual perception-oriented color model that contains three channels, i.e., hue, saturation and value, corresponding respectively to the color, the color depth, and the degree of brightness. HSV color space has two major advantages. The first advantage is that the value channel and the color channels are independent of each other; the other advantage is that the hue channel and the saturation channel are more suitable for human perception of color. CIELab, another visual perception-oriented color model, is more uniform in a visual sense. Euclidean distance can be used to measure the dissimilarity between colors in the CIELab color space.

According to the above advantages of the two color spaces, we combine the H channel of HSV color space with the L, a, and b channels of the CIELab color space to build a color feature map for objects' saliency computation. First, the original remote sensing image is converted from the RGB color space to the HSV color space and the CIELab color space. Second, values in the H, L, a, and b channels are quantized into 4, 8, 16, and 16 levels, respectively. Third, these four channels are combined into an 8192 ($4 \times 8 \times 16 \times 16$) level color feature map. Finally, the color histogram of all pixels within each segmented region of the image is calculated to obtain the objects' color features.

(2) Intensity Feature Extraction

As one of the three core features of human perception of color, the intensity feature has attracted wide attention and has been widely used in visual saliency computation. In this paper, we use a similar method of extracting color features to acquire objects' intensity features in images. First, the original remote sensing image is converted from RGB space to HSV space. Second, values in the V channel are quantized to 256 levels. Further, the intensity histogram of all pixels within each segmented region of the image is calculated to obtain the intensity feature of each object. Assuming that the original remote sensing image is I, and the three color channels of the RGB color space are R, G, and B, respectively, the corresponding intensity feature map V can be easily calculated by the following equation:

$$V = \frac{1}{3}(R + G + B) \tag{1}$$

(3) Texture Feature Extraction

As a type of commonly used low-level visual features, texture features have attracted more and more attention in the VAM field in recent years [29], [30]. In this study, we employ the rotation-invariant local binary pattern (LBP) [31] to compute the objects' visual saliency.

The texture feature extraction process for objects in the image is similar to the color feature extraction process. First, the original remote sensing image is converted into a grayscale image, from which the LBP feature map and local contrast (LC) feature map (quantized to 8 levels) are extracted. Then, the LBP feature map and the LC feature map are combined into a texture feature map. Since the rotation-invariant LBP pattern has only 36 possible values, LC and LBP feature maps can be combined as a 288 (36×8) level texture feature map. Finally, the texture histogram of all pixels within each segmented region of the image is calculated to obtain the texture feature of each object. Note that the texture feature represents the joint probability distribution of the LBP values and the LC values.

For all the segmented regions of the image, after extracting the above three types of visual features, we construct a complete object set, serving as a state space of the Markov chain in the next step. Given a remote sensing image I, SEG_n is the segmentation result of image I under scale n, where $n = 1, 2, ..., N$, N is the number of scale levels, $r_i^{(n)}$ is a segmented region in SEG_n, where $i = 1, 2, ..., R(n)$, $R(n)$ is the total number of the segmented regions under scale n. To construct object $Obj_i^{(n)}$ based on the segmented region $r_i^{(n)}$, we extract the following properties and visual features of the region $r_i^{(n)}$:

① Area $area_i^{(n)}$, i.e., the total number of pixels within the region $r_i^{(n)}$.

② Center coordinates $center_i^{(n)} = (x_i^{(n)}, y_i^{(n)})$, i.e., the mean 2D coordinates of all pixels within the region $r_i^{(n)}$.

③ Color feature vector $Clr_i^{(n)} = (H_1^{Clr}, H_2^{Clr}, ..., H_{8192}^{Clr})$, i.e., color histogram of the region $r_i^{(n)}$.

④ Intensity feature vector $Int_i^{(n)} = (H_1^{Int}, H_2^{Int}, ..., H_{256}^{Int})$, i.e., intensity histogram of the region $r_i^{(n)}$.

⑤ Texture feature vector $Tex_i^{(n)} = (H_1^{Tex}, H_2^{Tex}, ..., H_{288}^{Tex})$, i.e., texture histogram of the region $r_i^{(n)}$.

⑥ Set of adjacent objects $\{Obj_k^{(n)} \mid k \in \mathrm{NB}(Obj_i^{(n)})\}$, where $\mathrm{NB}(Obj_i^{(n)})$ records all index numbers of the adjacent objects of $Obj_i^{(n)}$.

With these properties and visual features of segmented regions in SEG_n, we further construct an object set $\{Obj_i^{(n)}\}_{i=1}^{R(n)}$ under scale n.

C. FOA Transition Probability Computation

Studies on the human visual cortex show that the receptive fields of most neurons present as concentric circles, while the center neurons and the surrounding neurons are in a mutual inhibition competition. Based on this evidence, Schiller et al. suggested that the mammalian visual system has both an ON channel (with a central activated region and a surrounding

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3179461, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

inhibited region) and an OFF channel (with a central inhibited region and a surrounding activated region) to yield equal sensitivity and to facilitate high contrast sensitivity [32].

Inspired by this study, we hypothesize that the saliency of an individual object is primarily determined by the feature contrast between itself and its adjacent objects. Based on this assumption, we further hypothesize that the edge weight between objects is mainly determined by two factors, i.e., the visual feature differences and the centroid distances between adjacent objects. The calculation of edge weight (taking the color feature as an example) follows:

$$\begin{cases} w_{i,k}^{(n)} = D_{Fea}(Obj_i^{(n)}, Obj_k^{(n)}) \cdot D_{Spt}(Obj_i^{(n)}, Obj_k^{(n)}) \\ D_{Fea}(Obj_i^{(n)}, Obj_k^{(n)}) = \exp(N(\chi^2(Clr_i^{(n)}, Clr_k^{(n)}))) - 1 \\ D_{Spt}(Obj_i^{(n)}, Obj_k^{(n)}) = \exp(-(N(\left\| center_i^{(n)} - center_k^{(n)} \right\|))^2 / c_1) \end{cases} \quad (2)$$

where $w_{i,k}^{(n)}$ is the edge weight between $Obj_i^{(n)}$ and its adjacent objects $Obj_k^{(n)}$. If two objects are not adjacent in the image, the edge weight between them should be zero. $D_{Fea}(Obj_i^{(n)}, Obj_k^{(n)})$ represents the visual feature differences between the two objects, and it can be replaced by the corresponding feature vector when calculating the intensity or texture feature difference instead. $D_{Spt}(Obj_i^{(n)}, Obj_k^{(n)})$ represents the centroid distances between the objects. Moreover, $\chi^2(Clr_i^{(n)}, Clr_k^{(n)})$ represents the chi-square distance between the color feature vectors, $Clr_i^{(n)}$ and $Clr_k^{(n)}$. For the feature vectors $A=(a_1, a_2, ..., a_T)$ and $B=(b_1, b_2, ..., b_T)$, their chi-square distance can be calculated as follows [33]:

$$\chi^2(A,B) = \sum_{t=1}^{T} \frac{(a_t - b_t)^2}{2(a_t + b_t)} \quad (3)$$

In addition, $\left\| center_i^{(n)} - center_k^{(n)} \right\|$ represents the Euclidean distance between the centroids of the two objects. $N(\cdot)$ is a linear normalization function:

$$N(m) = \frac{m}{\max(M)} \quad (4)$$

where m is an arbitrary element of matrix M and c_1 is a constant parameter (its settings are discussed in a sensitivity analysis).

After sderving all edge weights between objects, the transition probability $p_{i,k}^{(n)}$ of FOA between $Obj_i^{(n)}$ and its adjacent object $Obj_k^{(n)}$ can be calculated as follows:

$$p_{i,k}^{(n)} = w_{i,k}^{(n)} \Big/ \sum_{k=1}^{K} w_{i,k}^{(n)} \quad (5)$$

D. Visual Saliency Computation

After computing the FOA transition probabilities between all objects, an FOA transition probability matrix between objects can be built. Further, the FOA equilibrium distribution among objects can be calculated. Assuming that the FOA equilibrium distribution is $\Pi^{(n)} = (\pi_1^{(n)}, \pi_2^{(n)}, ..., \pi_{R(n)}^{(n)})$, and the transition probability matrix is $P^{(n)}$:

$$P^{(n)} = \begin{pmatrix} p_{1,1}^{(n)} & \cdots & p_{1,k}^{(n)} & \cdots & p_{1,R(n)}^{(n)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{R(n),1}^{(n)} & \cdots & p_{R(n),k}^{(n)} & \cdots & p_{R(n),R(n)}^{(n)} \end{pmatrix} \quad (6)$$

According to the nature of equilibrium distribution, the relationship between $\Pi^{(n)}$ and $P^{(n)}$ conforms to the following formula:

$$\Pi^{(n)} = \Pi^{(n)} \cdot P^{(n)} \quad (7)$$

In practice, any element $\pi_i^{(n)}$ of the equilibrium distribution $\Pi^{(n)}$ can be quickly calculated by the edge weights $w_{i,k}^{(n)}$ from Equation (2), which is calculated as [34]:

$$\pi_i^{(n)} = \sum_k w_{i,k}^{(n)} \Big/ \sum_{i,k} w_{i,k}^{(n)} \quad (8)$$

where $\sum_k w_{i,k}^{(n)}$ is the sum of the edge weights between $Obj_i^{(n)}$ and all its adjacent objects, $\sum_{i,k} w_{i,k}^{(n)}$ is the sum of all edge weights in the graph.

Besides the FOA equilibrium distribution, we argue that the visual saliency of an object is also closely associated with the area of the object. For a remote sensing image with a relatively stable segmentation result, objects with very large areas tend to be the background of the image. Thus, when calculating the area factors of an object, its visual saliency with very large areas should be inhibited. The formula of the area factor can be expressed as

$$A_i^{(n)} = 1 / (1 + (c_2 \cdot area_i^{(n)} / (iw \cdot ih))^2) \quad (9)$$

where iw and ih are the width and height of the original remote sensing image, respectively, c_2 represents a constant parameter (discussed in the sensitivity analysis). After obtaining the equilibrium distribution $\pi_i^{(n)}$ and the area factor $A_i^{(n)}$ of each object by Equation (8) and (9), respectively, visual saliency $S_i^{(n)}$ of the object $Obj_i^{(n)}$ can be calculated as

$$S_i^{(n)} = \pi_i^{(n)} \cdot A_i^{(n)} \quad (10)$$

To generate a normalized saliency map under each segmentation scale, we normalize the visual saliencies of the objects:

$$Saliency_i^{(n)} = \frac{S_i^{(n)} - \min(\{S_i^{(n)}\})}{\max(\{S_i^{(n)}\}) - \min(\{S_i^{(n)}\})} \quad (11)$$

By assigning the normalized visual saliency values to all pixels in the corresponding objects, we can obtain the scale – wise normalized saliency map.

E. Saliency Map Fusion

After obtaining the normalized visual saliency maps under multiple scales, we further fuse them into a single visual saliency map. Assuming that pxl is a pixel in the original remote sensing image I, its multi-scale saliency can be expressed as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3179461, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

$$SMap(pxl)=\frac{\sum_{n=1}^{N}\sum_{i=1}^{R(n)} Saliency_i^{(n)}\cdot\left(\| I_{pxl}-c_i^{(n)}\|+\varepsilon\right)^{-1}\cdot l(pxl\in Obj_i^{(n)})}{\sum_{n=1}^{N}\sum_{i=1}^{R(n)}\left(\| I_{pxl}-c_i^{(n)}\|+\varepsilon\right)^{-1}\cdot l(pxl\in Obj_i^{(n)})} \tag{12}$$

where I_{pxl} is the three-dimensional feature vector in RGB color space of the pixel pxl, $c_i^{(n)}$ is the color center of all the pixels in $Obj_i^{(n)}$, ε is a small constant (set as 0.1 in our experiment), and $l(pxl\in Obj_i^{(n)})$ is an indicator function, whose specific values are as follows:

$$l(pxl\in Obj_i^{(n)})=\begin{cases}1, & pxl\in Obj_i^{(n)}\\ 0, & pxl\notin Obj_i^{(n)}\end{cases} \tag{13}$$

The multi-scale visual saliency of each pixel can be obtained by Equation (12). We further normalize the visual saliencies of all pixels again to obtain the final saliency map of the original remote sensing image.

IV. EXPERIMENTAL ANALYSIS

A. Experiment Dataset

UCM dataset: The UCM dataset [35] consists of 21 image categories, and each category has 100 images with the size of 256×256 pixels. We selected 600 images containing distinct ground objects from 8 categories, which are airplane (100 images), baseball diamond (100 images), freeway (59 images), golf course (65 images), river (67 images), sparse residential (58 images), storage tanks (97 images), and tennis courts (54 images). These images are then manually labeled to generate the ground truth masks for performance evaluation.

ORSSD dataset: The ORSSD dataset [36] is a challenging dataset with diverse spatial resolutions including 1264×987, 800×600, and 256×256. It contains 800 optical remote sensing images collected from several existing datasets.

B. Evaluation Measures

We employ two sets of evaluation criteria to compare the performances of different methods quantitatively, i.e., the precision-recall curve (i.e., PR curve) with the F-measure curve for full-range thresholds and the average precision, recall, and F-measure for adaptive thresholds.

The PR curve and the F-measure curve for full-range thresholds are calculated as follows. First, an integer value within the range [0,255] is selected as a threshold for generating a binary mask from the saliency map. Second, the precision and recall can be calculated by comparing the binary mask B and the ground truth mask G:

$$\begin{cases}Precision=\dfrac{|B\cap G|}{|B|}\\ \\ Recall=\dfrac{|B\cap G|}{|G|}\end{cases} \tag{14}$$

where $|\cdot|$ denotes the number of non-zero entries in the mask. The F-measure can be calculated as

$$F-measure=\frac{(1+\beta^2)\cdot Precision\cdot Recall}{\beta^2\cdot Precision+Recall} \tag{15}$$

where β^2 is a weight parameter, commonly set to 0.3 to increase the importance of the precision value [14], [18], [39]. Third, for each integer threshold within the range [0,255], the precision, recall and F-measure are calculated based on the binary mask and the ground truth mask. Fourth, for all images in the experimental dataset, the precision, recall and F-measure values for full-range thresholds are calculated, and then the average precision, recall and F-measure for each threshold within the range [0,255] can be obtained. Finally, the PR curve is plotted by setting the average precision and average recall as the ordinate and abscissa values, respectively; the F-measure curve is plotted by setting the average F-measure and the threshold as the ordinate and abscissa values, respectively.

The average precision, recall, and F-measure for adaptive thresholds are calculated as follows. First, the Otsu method [39] is employed to perform an adaptive binarization on the saliency map. Second, the precision, recall and F-measure are calculated via Equations (14) and (15) based on the adaptive binary mask and the ground truth mask. Finally, for all images in the experimental dataset, the precision, recall and F-measure values are calculated, and then the average precision, recall and F-measure for adaptive thresholds can be obtained.

C. Performance Analysis

(1) Visual saliency computation based on different segmentation methods

As a key step in our ORWVS model, image segmentation has a great impact on the subsequent object extraction and visual saliency computing. To analyze the performances of different segmentation methods on visual saliency computation, we employ the GS, QS and SLIC segmentation methods and their two combination schemes to segment the original remote sensing images at multiple scales and calculate the final saliency map. The first scheme is to compute the mean values of the saliency maps based on the three methods at the pixel level (named "Mean"), and the other scheme is to compute the max values of the three saliency maps at the pixel level (named "Max").

The segmentation results measured using average precision, recall, and F-measure for adaptive thresholds are reported in Table I. From Table I, we notice that, in terms of average precision and F-measure, GS outperforms the other two methods. However, when using the average recall as an evaluation measure, QS and SLIC perform notably better than GS, while QS achieves the best performance among the three methods. Such inconsistent rankings are mainly due to the adaptive thresholds. Our further analysis suggests that the adaptive threshold values of the QS and SLIC saliency maps are less than that of GS saliency maps in general. Therefore, QS and SLIC saliency maps often have a larger foreground area, leading to a higher recall and a lower precision. However, according to Equation (15), precision is more important for the F-measure calculation. Thus, the performance rankings using

average F-measure are similar to that using average precision and different from that using average recall.

TABLE I
PERFORMANCES OF THE GS, SLIC, QS SEGMENTATION METHODS AND THEIR FUSIONS ON AVERAGE PRECISION, RECALL, AND F-MEASURE FOR ADAPTIVE THRESHOLDS

Datasets	Methods	Evaluation measures		
		Precision	Recall	F-measure
UCM	GS	***0.6564***	0.7555	***0.6370***
	SLIC	0.4869	0.6211	0.4761
	QS	0.4722	0.6372	0.4688
	Mean	0.5800	0.7546	0.5826
	Max	0.5030	***0.7982***	0.5229
ORSSD	GS	***0.4739***	0.6884	***0.4770***
	SLIC	0.4408	0.6616	0.4476
	QS	0.4715	0.6390	0.4674
	Mean	0.4543	0.6685	0.4639
	Max	0.3890	***0.7020***	0.4412

(2) Visual saliency computation based on different object features

Object feature extraction is another critical step of our proposed ORWVS model. To thoroughly analyze the performances of different object features on visual saliency computation, we extract color, intensity, and texture features of the objects to construct the image object sets and then compute the visual saliency for objects to obtain their corresponding visual saliency maps.

The performances of different object features on salient object extraction are shown in Table II. With regard to the average recall and F-measure, the color feature outperforms the other two features, particularly the texture feature. However, as for the average precision measure, the performance rankings of these three features follow intensity, color and texture, with intensity and color performing significantly better than the texture. Similar to the experimental results presented in Table I, due to the greater importance of precision in the calculation of F-measure, the performance rankings using average F-measure are similar to that using average precision and different from that using average recall.

TABLE II
PERFORMANCES OF THE COLOR, INTENSITY, TEXTURE FEATURES AND THEIR COMBINATIONS ON AVERAGE PRECISION, RECALL, AND F-MEASURE FOR ADAPTIVE THRESHOLDS

Datasets	Methods	Evaluation measures		
		Precision	Recall	F-measure
UCM	Color	0.6564	0.7555	0.6370
	Intensity	***0.6636***	0.7233	0.6338
	Texture	0.5889	0.5904	0.5236
	Mean	0.6531	0.7631	***0.6380***
	Max	0.6106	***0.7849***	0.6110
ORRSD	Color	0.4739	0.6884	***0.4770***
	Intensity	***0.4742***	0.6713	0.4748
	Texture	0.4512	0.6361	0.4518
	Mean	0.4711	0.7075	0.4720
	Max	0.4306	***0.7192***	0.4449

(3) Sensitivity analysis to model parameters

There are two important hyperparameters in our proposed ORWVS model, i.e., c_1 and c_2. We evaluate the performance with different values of c_1 (Table III) and c_2 (Table IV).

TABLE III
PERFORMANCES OF THE ORWVS MODEL WITH DIFFERENT VALUE OF C_1 ON AVERAGE PRECISION, RECALL, AND F-MEASURE FOR ADAPTIVE THRESHOLDS

Datasets	C_1 value	Evaluation measures		
		Precision	Recall	F-measure
UCM	0.2	0.6543	0.6603	0.6088
	0.4	***0.6619***	0.7258	0.6351
	0.6	0.6584	0.7431	0.6360
	0.8	0.6564	0.7555	***0.6370***
	1.0	0.6489	***0.7596***	0.6321
ORRSD	0.2	***0.4756***	0.6567	0.4701
	0.4	0.4739	0.6884	***0.4770***
	0.6	0.4655	0.7198	0.4703
	0.8	0.4620	***0.7318***	0.4677
	1.0	0.4614	0.7271	0.4669

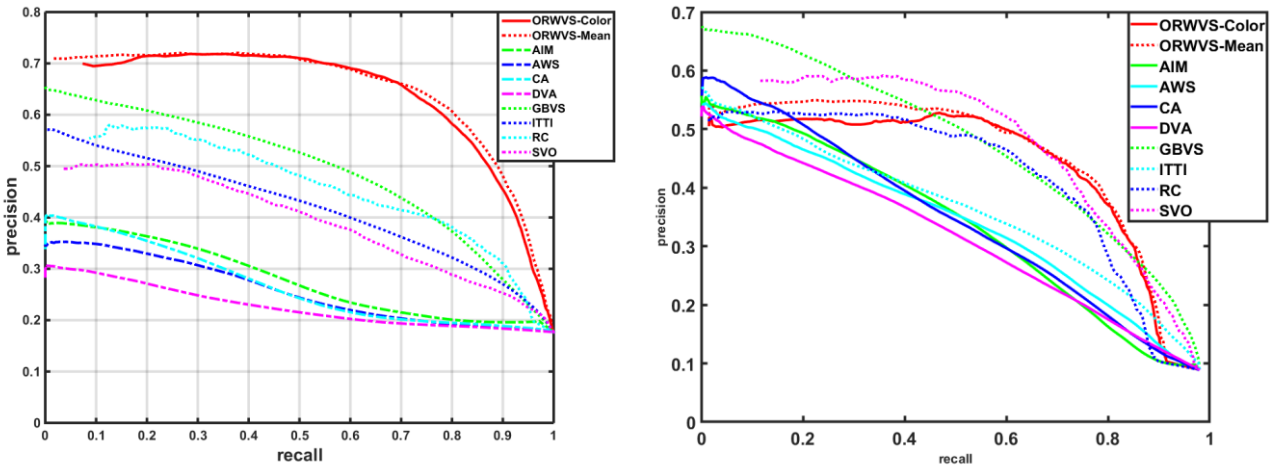
TABLE IV
PERFORMANCES OF THE ORWVS MODEL WITH DIFFERENT VALUE OF C_2 ON AVERAGE PRECISION, RECALL, AND F-MEASURE FOR ADAPTIVE THRESHOLDS

Datasets	C_2 value	Evaluation measure		
		Precision	Recall	F-measure
UCM	0.5	0.6491	***0.7592***	0.6303
	0.6	0.6493	0.7575	0.6306
	0.7	0.6497	0.7574	0.6313
	0.8	0.6527	0.7561	0.6334
	0.9	***0.6564***	0.7555	***0.6370***
	1.0	***0.6564***	0.7537	0.6368
ORRSD	***0.4740***	0.6926	0.4765	***0.4740***
	0.4737	0.6929	0.4762	0.4737
	0.4740	0.6944	0.4769	***0.4740***
	0.4733	***0.6954***	0.4761	0.4733
	0.4733	0.6902	0.4762	0.4733
	0.4739	0.6884	***0.4770***	0.4739

(4) Comparison with other VAMs

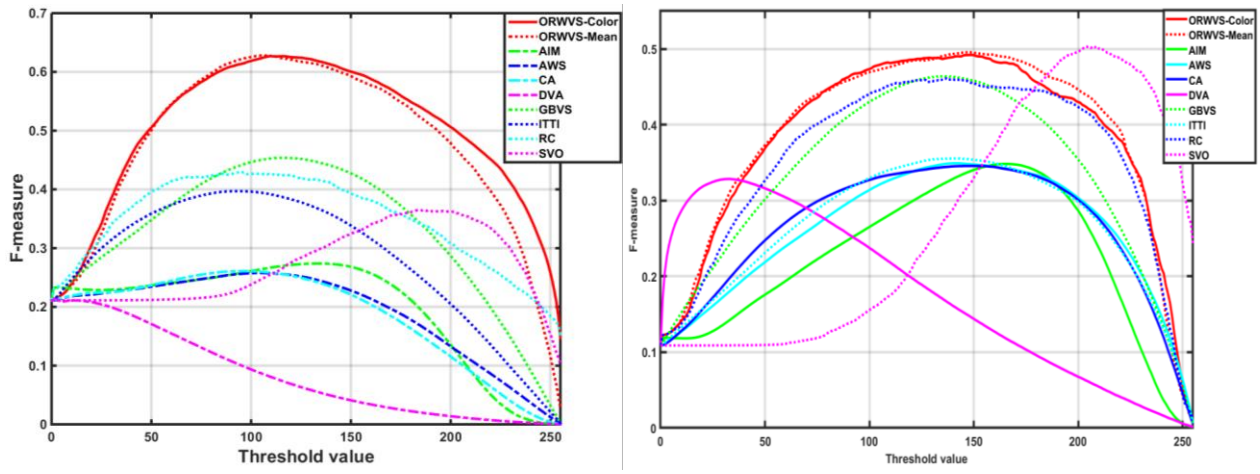
To validate the performance and advantage of our ORWVS model, we compare this model with other eight VAMs. These eight models are AIM [10], AWS [12], CA [13], DVA [11], GBVS [4], ITTI [3], RC [37], and SVO [39]. We conduct a detailed comparative analysis between our ORWVS model and the other eight models. The PR curves and F-measure curves of all these models are presented in Fig. 3 and 4, respectively.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3179461, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing



(a) Results on UCM dataset. (b) Results on ORSSD dataset.

Fig. 3. PR curves of the ORWVS model and the other eight models on two datasets.



(a) Results on UCM dataset. (b) Results on ORSSD dataset.

Fig. 4. F-measure curves of the ORWVS model and the other eight models on two datasets.

To compare the performances of our ORWVS model and the other eight models on salient object extraction, we further calculate the average precision, recall, and F-measure for adaptive thresholds. The results are documented in Table V. We notice that our ORWVS model outperforms the other eight models by a large margin in terms of average precision and F-measure on both UCM and ORSSD datasets. However, the GBVS model performs the best in terms of average recall. In fact, pixel-based models generally perform well on average recall, mainly due to the Gaussian smoothing process for saliency map generation in these pixel-based models, which brings in a high recall as well as a low precision. However, for salient object extraction, precision is considered more important than recall. Thus, in our evaluation, we value precision more than recall.

TABLE V
PERFORMANCES OF THE COLOR, INTENSITY AND TEXTURE FEATURE ON AVERAGE PRECISION, RECALL, AND F-MEASURE FOR ADAPTIVE THRESHOLDS ON TWO DATASETS.

Datasets	VAM	Evaluation measure			Time(s)
		Precision	Recall	F-measure	
UCM	ORWVS-Color	***0.6564***	0.7555	0.6370	0.396
	ORWVS-Mean	0.6531	0.7631	***0.6380***	1.406
	AIM	0.2337	0.6309	0.2542	1.896
	AWS	0.2513	0.4853	0.2536	1.887
	CA	0.2752	0.4314	0.2637	64.028
	DVA	0.2799	0.1644	0.1828	1.072
	GBVS	0.3948	***0.7800***	0.4197	1.714

	ITTI	0.3840	0.6547	0.3958	***0.285***
	RC	0.4627	0.6285	0.4539	1.238
	SVO	0.3080	0.7435	0.3299	147.393
ORRSD	ORWVS-Color	***0.6564***	0.7555	0.6370	0.396
	ORWVS-Mean	0.6531	0.7631	***0.6380***	1.406
	AIM	0.2337	0.6309	0.2542	1.896
	AWS	0.2513	0.4853	0.2536	1.887
	CA	0.2752	0.4314	0.2637	64.028
	DVA	0.2799	0.1644	0.1828	1.072
	GBVS	0.3948	***0.7800***	0.4197	1.714
	ITTI	0.3840	0.6547	0.3958	***0.285***
	RC	0.4627	0.6285	0.4539	1.238
	SVO	0.3080	0.7435	0.3299	147.393

For visual analysis, we take UCM dataset as an example and present its visual saliency maps by different VAMs. The results are illustrated in Fig.5. It can be observed that our model and particularly ORWVS-Color outperforms other competing models.

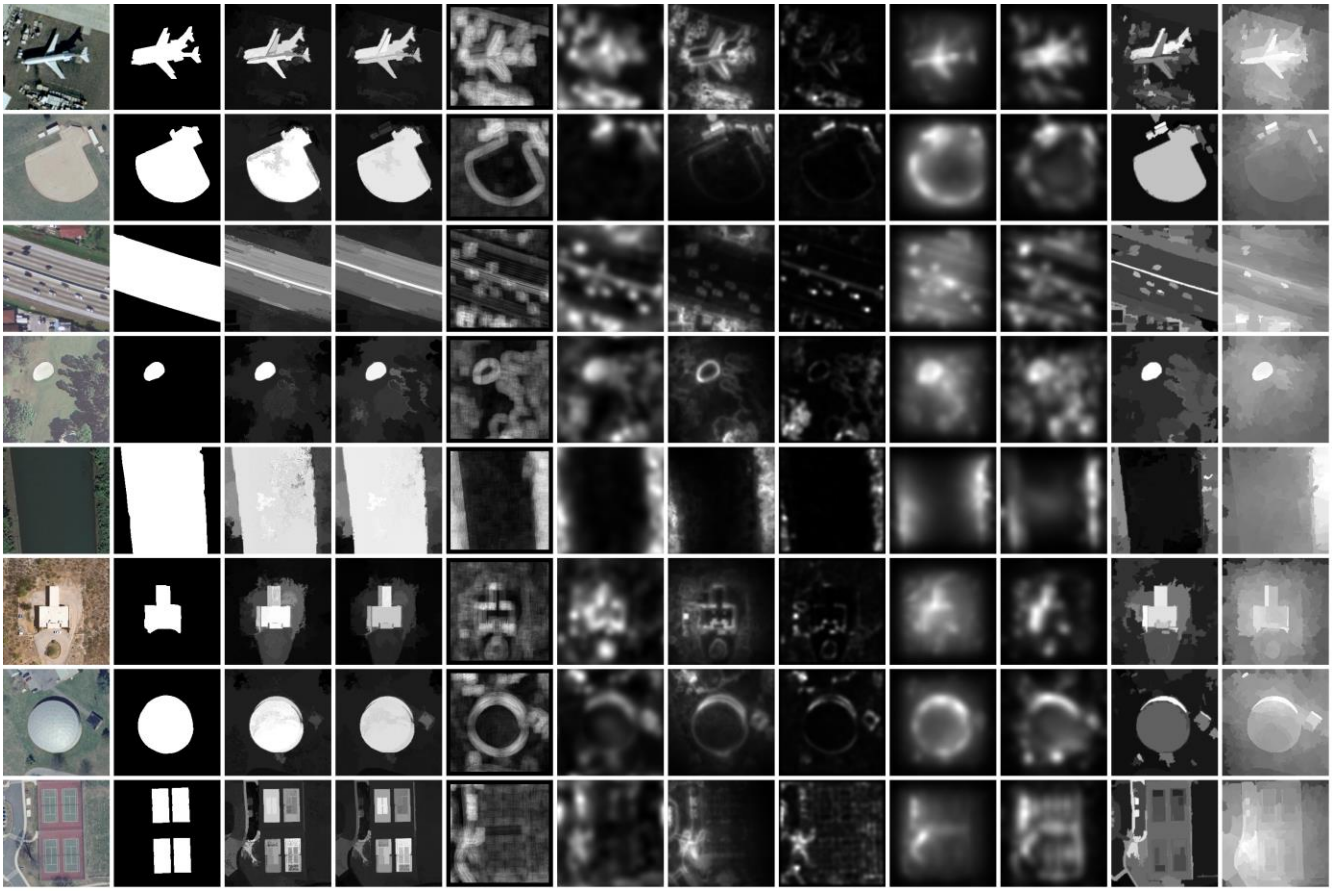


Fig. 5. Comparison of the visual saliency maps of UCM by different VAMs (1st column are sample images, 2nd column are ground truth mask images, 3rd-12th columns are saliency maps by the ORWVS-Color, ORWVS-Mean, AIM, AWS, CA, DVA, GBVS, ITTI, RC and SVO models, respectively.)

V. Conclusion

The merging remote sensing platforms that provide high-resolution images provide great opportunities as well as challenges. We notice that the efficient and accurate feature extraction for ground objects in images has attracted increasing attention. In this study, we introduce the selective attention mechanism of HVS into remote sensing image processing and

employ the visual attention models to extract ROI (i.e., salient region) from remote sensing images. We propose a new visual attention model, i.e., the object-oriented random walk model for visual saliency detection (ORWVS). The proposed views image objects as the basic units in visual saliency computation, leading to its great computation efficiency and accuracy in terms of salient object extraction. We further analyze the performances of different segmentation methods and different object features in the model and conduct a comprehensive comparative analysis between our ORWVS model and other eight VAMs. Experimental results show that, in the saliency maps generated by our ORWVS model, salient objects own clear edges and accurate contours. Quantitative evaluations suggest that the overall performance of our proposed model is superior to that of the other eight VAMs.

The ORWVS model can be widely used for object detection and recognition, change detection, image retrieval, and scene understanding from high-resolution remote sensing images. As for future works, we plan to further improve this model in two directions, i.e., the efficiency of the multi-scale image segmentation method and the development of innovative object features.

ACKNOWLEDGMENTS

We give sincere thanks to Dr. Xiao Huang and the anonymous reviewers for their valuable suggestions, who provided great help in improving our manuscript.

REFERENCES

[1] D. Li, Q. Tong, R. Li, J. Gong and L. Zhang, "Current issues in high-resolution earth observation technology," *Science China-Earth Sciences*, vol. 55, no. 7, pp. 1043-1051, Jul. 2012.

[2] M. Li, T. Pan, H. Guan, H. Liu, and J. Gao, "Gaofen-2 mission introduction and characteristics," *Proceedings of the 66th International Astronautical Congress (IAC 2015)*, Jerusalem, Israel, paper: IAC-15-B1.2.7, Oct. 2015.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[4] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *in Advances in Neural Information Processing Systems 19*, Vancouver, Canada, pp. 545-552, 2006.

[5] A. Borji, and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185-207, Jan. 2013.

[6] A. M. Treisman, and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97-136, Jan. 1980.

[7] C. Koch, and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[8] D. E. Rumelhart, and D. Zipser, "Feature discovery by competitive learning," *Cognitive Science*, vol. 9, no. 1, pp. 75-112, 1985.

[9] H. Bouma, and D.G. Bouwhuis, *Attention and Performance*. Hillsdale, N.J.: Erlbaum, vol. 10, pp. 531–556, 1984.

[10] B. Bruce, and J. K. Tsotsos, "Saliency based on information maximization," *in Advances in Neural Information Processing Systems 18*, Vancouver, Canada, pp. 155-162, 2005.

[11] X. Hou, and L. Zhang, "Dynamic visual attention: Searching for coding length increments", in Advances in Neural Information Processing Systems 21, Vancouver, Canada, pp. 681-688, 2008.

[12] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Decorrelation and distinctiveness provide with human-like saliency," *in Proceedings of the Advanced Concepts for Intelligent Vision Systems (ACIVS 2009)*, pp. 343-354. 2009.

[13] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915-1926, Oct. 2012.

[14] A. Borji, M. M. Cheng, H. Z. Jiang, and J. Li, "Salient object detection: a benchmark," *IEEE Transactions on Image Processing*, vol. 24, no.12, pp. 5706-5722, Dec. 2015.

[15] D. Brockmann, and T. Geisel, "Are human scanpaths Levy flights?," *in Proceedings of the 9th International Conference on Artificial Neural Networks*, Edinburgh, Scotland, vol. 1&2, no. 470, pp. 263-268, 1999.

[16] G. Boccignone, and M. Ferraro, "Modelling gaze shift as a constrained random walk," Physica A, vol. 331, no. 1-2, pp. 207-218, Jan. 2004.

[17] Y. Pang, X. Yu, Y. Wu, et al. "Bagging-based saliency distribution learning for visual saliency detection," *Signal Processing: Image Communication*, vol. 87, pp. 115928, 2020.

[18] R. Cong, J. Lei, H. Fu, et al. "Review of visual saliency detection with comprehensive information," *IEEE Transactions on circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2941-2959, 2018.

[19] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: a benchmark", *in Proceedings of the 12th European Conference on Computer Vision (ECCV)*, Florence, Italy, vol. LNCS7573, pp. 414-429, Oct. 2012.

[20] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 65, no.1, pp. 2-16, Jan. 2010.

[21] R. L. Kettig, and D. A. Landgrebe. "Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects," *IEEE Transactions on Geoscience Electronics*, vol. GE-14, no. 1, pp. 19-26, Jan. 1976.

[22] S. Georganos, T. Grippa, S. Vanhuysse, et al. "Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application," *GIScience & remote sensing*, vol. 55, no. 2, pp. 221-242, 2018.

[23] C. Zhang, G. Li, W. Cui. "High-resolution remote sensing image change detection by statistical-object-based method," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 7, pp. 2440-2447, 2018.

[24] J. Wu, B. Li, Y. Qin, et al. "An object-based graph model for unsupervised change detection in high resolution remote sensing images," *International Journal of Remote Sensing*, vol. 42, no.16, pp. 6209-6227, 2021.

[25] L. G. Shapiro, and G. C. Stockman. *Computer Vision*. New Jersey: Prentice-Hall, pp. 279-325, 2001.

[26] R. Achanta, A. Shaji, K. Smith K, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274-2282, Nov. 2012.

[27] P. F. Felzenszwalb, and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167-181, Sep. 2004.

[28] A Vedaldi, and S. Soatto, "Quick shift and kernel methods for mode seeking," *in Proceedings of the 10th European Conference on Computer Vision (ECCV)*, Marseille, France, vol. LNCS5305, pp. 705-718, Oct. 2008.

[29] J. Chi, C. Wu, X. Yu, et al. "Saliency detection via integrating deep learning architecture and low-level features," *Neurocomputing*, vol. 352, pp. 75-92, 2019.

[30] L. Zhang, Y. Liu. "Remote Sensing Image Generation Based on Attention Mechanism and VAE-MSGAN for ROI Extraction," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2021.

[31] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no.7, pp. 971-987, Jul. 2002.

[32] P. H. Schiller, J. H. Sandell, and J. H. R. Maunsell, "Functions of the ON and OFF channels of the visual system," *Nature*, vol. 322, pp. 824-825, Aug. 1986.

[33] Y. Rubner, and C. Tomasi, "*Perceptual metrics for image database navigation*," New York: Springer US, 2001.

[34] W. Wang, Y. Z. Wang, Q. M. Huang, and W. Gao. "Measuring Visual Saliency by Site Entropy Rate," *in Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2368-2375, 2010.

[35] Y. Yang, and S. Newsam. "Bag-of-visual-words and spatial extensions for land-use classification," *in Proceedings of the 18th ACM SIGSPATIAL*

International Conference on Advances in Geographic Information Systems, San Jose, United States, pp. 270-279, 2010.

[36] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian and S. Kwong, "Nested Network With Two-Stream Pyramid for Salient Object Detection in Optical Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9156-9166, Nov. 2019.

[37] S. Bhuyan, D. Sen, S. Deb. "Structure-aware multiple salient region detection and localization for autonomous robotic manipulation," *IET Image Proce*ssing, pp. 1-27, Jan. 2022.

[38] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-9, no. 1, pp. 62-66, Jan. 1979.

[39] K. Y. Chang, T. L. Liu, H. T. Chen, and S. H. Lai. "Fusing generic objectness and visual saliency for salient object detection," *in Proceedings of the 13th International Conference on Computer Vision (ICCV)*, pp. 914-921, 2011.

Lin Ding received the B.S. degree from Wuhan University, Wuhan, China, in 2013. He is working toward the Ph.D. degree with the School of Remote Sensing and Information Engineering, Wuhan University, China.

His research interests include urban impervious surface extraction and environment monitoring.

Xing Wang received the B.S. degree from Wuhan University, Wuhan, China, in 2008 and the Ph.D. degree in photogrammetry and remote sensing from the School of Remote Sensing and Information Engineering, Wuhan University, in 2015.

He is currently a Lecturer with the School of Marine Science and Technology, Tianjin University. His research interests include remote sensing image processing, object detection and image retrieval.

Deren Li received the bachelor's and master's degrees from Wuhan University, Wuhan, China, and the Ph.D. degree from University of Stuttgart, Stuttgart, Germany, in 1963, 1981, and 1985, respectively, all in photogrammetry and remote sensing.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China.

Prof. Li was selected as a member of Chinese Academy of Sciences in 1991 and a member of Chinese Academy of Engineering in 1994. He was awarded the title of honorary doctor from ETH Zürich, Switzerland in 2008.