ARTICLE IN PRESS
Chinese Journal of Aeronautics, (2019), xxx(xx): xxx–xxx

Chinese Society of Aeronautics and Astronautics
& Beihang University
Chinese Journal of Aeronautics
cja@buaa.edu.cn
www.sciencedirect.com

A novel visual attention method for target detection from SAR images

Fei GAO [a], **Aidong LIU** [a], **Kai LIU** [a,*], **Erfu YANG** [b], **Amir HUSSAIN** [c]

[a] *School of Electronic and Information Engineering, Beihang University, Beijing 100083, China*
[b] *Space Mechatronic Systems Technology Laboratory, Department of Design, Manufacture and Engineering Management, University of Strathclyde, Glasgow G1 1XJ, UK*
[c] *Cognitive Big Data and Cyber-Informatics (CogBID) Laboratory, School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, Scotland, UK*

Received 3 June 2018; revised 23 November 2018; accepted 14 February 2019

KEYWORDS

Learning strategy;
Synthetic Aperture Radar (SAR) images;
Target detection;
Top-down;
Visual attention mechanism

Abstract Synthetic Aperture Radar (SAR) imaging systems have been widely used in civil and military fields due to their all-weather and all-day abilities and various other advantages. However, due to image data exponentially increasing, there is a need for novel automatic target detection and recognition technologies. In recent years, the visual attention mechanism in the visual system has helped humans effectively deal with complex visual signals. In particular, biologically inspired top-down attention models have garnered much attention recently. This paper presents a visual attention model for SAR target detection, comprising a bottom-up stage and top-down process. In the bottom-up step, the Itti model is improved based on the difference between SAR and optical images. The top-down step fully utilizes prior information to further detect targets. Extensive detection experiments carried out on the benchmark Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset show that, compared with typical visual models and other popular detection methods, our model has increased ability and robustness for SAR target detection, under a range of Signal to Clutter Ratio (SCR) conditions and scenes. In addition, results obtained using only the bottom-up stage are inferior to those of the proposed method, further demonstrating the effectiveness and rationality of a top-down strategy. In summary, our proposed visual attention method can be considered a potential benchmark resource for the SAR research community.

1. Introduction

The human visual system can identify multiple targets from complicated visual scenes, which surpasses the performance of state of the art computer vision technologies. It is believed that the biological visual system is a hierarchical structure, where stimuli initially activate the retina and are then transmitted to the optic chiasm via optic nerves – termed the low-level

* Corresponding author.
E-mail address: liuk@buaa.edu.cn (K. LIU).
Peer review under responsibility of Editorial Committee of CJA.

Production and hosting by Elsevier

https://doi.org/10.1016/j.cja.2019.03.021
1000-9361 © 2019 Chinese Society of Aeronautics and Astronautics. Production and hosting by Elsevier Ltd.

Please cite this article in press as: GAO F et al. A novel visual attention method for target detection from SAR images, *Chin J Aeronaut* (2019), https://doi.org/10.1016/j.cja.2019.03.021

stage. It then reaches a fork road, leading to two different pathways: a collicular pathway connected to the Superior Colliculus (SC), and a retino-geniculate pathway connected to the Lateral Geniculate Nucleus (LGN). From the LGN, the stimuli is transferred to the primary visual cortex (V1) and higher brain areas (V2–V4), etc., successively.[1] The aforementioned progress is termed the high-level stage, occurring primarily in the visual cortex of the brain.

Research has demonstrated that about 10^8–10^9 bits of visual data enter the visual system through the eyes every second,[2] which need effective processing mechanisms. Fortunately, within this system, there is an instant localization ability which guides attention to interesting areas. From a biological perspective, the retina contains two kinds of photoreceptors: rods and cones. The cones provide the color sensitivity and are few in number, whereas the rods are more numerous and provide luminance sensitivity. The fovea centralis lies in the center of the macula, where almost all rods are found, and a rod has a 0.3 mm diameter. The arrangement of these photoreceptors ensures that humans perceive only the currently focused part of a visual scene even when the targets appear very ion, whilst perceiving the rest in a coarse manner. This localization ability is termed visual attention, which helps humans effectively turn their attention to more salient areas in complex environments. In the biological theory, the saliency of objects is decided by the unique receptive fields of ganglion cells in the retina. The receptive fields can be segmented into a center and a surrounding area. For the same kind of stimuli, the reactions of two parts of the receptive fields are opposing. For example, if the center receptive field is activated by a stimulus, the surrounding receptive field is bound to be inhibited. The two actions offset each other. Hence, the signals from a flat area will not activate the receptive field and will be given low saliency.

Bottom-up cues: Inspired by the biological visual system, researchers are interested in modeling it into a mathematical computational system. Treisman and Gelade first proposed the Feature Integration Theory (FIT),[3] which established the basis for future attention models. Then a bottom-up model was proposed by Koch and Ullman,[4] introducing the saliency map[2] to model visual attention. Itti and Koch[5] proposed the first fully implemented bottom-up attention model in 2000. In their biologically inspired model, the input is decomposed into three features, i.e., intensity, color and orientation generated using three 9-scaled Gaussian-Pyramids. Following a center-surrounding operation, the raw feature maps are normalized into 6 intensity maps, 12 color maps, and 24 orientation maps, respectively. The feature maps are further combined with the final saliency maps. This model simulates the early stage of biological visual systems; hence the saliency map is in reasonable agreement with subjective human perceptions and has proven to be robust to noises.

However, this model ignores the important impact of acknowledgement, expectations and other prior information, for the moving of attention. Following this, many top-down models have been proposed, which consider both top-down and bottom-up cues.[2,6] In the following, some related works are further reviewed.

Top-down cues: It has been proven that the visual data processing of human beings exists as a top-down mechanism.[7,8] Unlike the bottom-up process, the top-down mechanism is slow, goal-driven, and prior-dependent. It is related to the higher areas in the brain. As the biological process of top-down attention has not been well understand in bioneurology, the number of top-down models is far fewer than bottom-up. Current top-down attention models have two subclasses. One utilizes prior knowledge (goal, etc.) to guide the fusion of low-level features. The first corresponding model was the revised guided search structure (GS2 model) raised in 1994.[9] This model and its following improved models direct attention towards interesting regions by changing the weights of various visual features.[10,11] Navalpakkam and Itti[12] took the signal-to-noise ratio as a cost function during the detection of targets, and proposed an optimal integration of top-down and bottom-up cues. In this method, the top-down cues utilize the statistical information of the targets and clutter, which maximizes target detection speed and is also sensitive to unexpected stimulus changes. The Visual Object detection with a CompUtational attention System (VOCUS) model is another top-down attention method for target detection proposed by Frintrop et al.,[1] which contains a search sub-model and a learning sub-model. VOCUS can reduce the impact of the illumination variances and viewpoint changes and has better detection performance, especially when many target classes are being considered. The top-down extension aims to find the best combination of weights from the training set and eventually generates a top-down saliency map. Kouchaki et al.[13] introduced the genetic algorithm in feature fusion, instead of a cross-scale and linear combination. This method has been proven to be effective in improving the speed of detection in a complex scene. In Ref. [14], Han et al. proposed that rough sets could be used to optimize the weights of each feature during the generation of the final saliency map. The results obtained from real data have demonstrated the effectiveness and generalization of this model. However, compared with the human brain, these works utilize limited prior information.

Other top-down attention models mainly utilize mathematical methods to simulate the top-down mechanism in the visual system, e.g. fuzzy theory, Conditional Random Field (CRF), etc. Tsotsos raised a new Winner-Takes-All (WTA) updating rule to model the competitive mechanism of the visual system,[15] hence the computational utility of the resulting method for robot vision is evident. The top-down model in Ref. [16] used a fuzzy adaptive resonance theory with the learning function to model visual memory. In order to obtain the top-down priors, Borji et al.[17] employed evolutionary algorithms to optimize the parameters of the basic saliency model. The comparative results against the basic saliency model indicate the merits of this method. Moreover, a Growing Fuzzy Topology Adaptive Resonance Theory (GFTART) was proposed in Ref. [18] containing two stages: calculating the bottom-up features of arbitrary targets and forming the top-down bias. GFTART is shown to perform well in perceiving the increasing arbitrary objects in real scenes, as well as effectively detecting the given targets. J.Yang and M.H. Yang[19] proposed a three-layered top-down visual saliency model, which from top to bottom consisted of a visual dictionary, sparse coding and a CRF. This model performs favorably against many other top-down attention models for target detection. However, those models are more mathematical, and not biologically plausible.

Application in SAR target detection of visual attention model: The SAR imaging system has all-weather and all-day abilities to detect and monitor the ground and sea. It is not

affected by climate and illumination, and is widely used in military and civil fields.[20] In particular, the development of recent SAR technology has vastly enhanced the resolution of images, making it difficult to process and analyze them. Consequently, the utilization of SAR images is only about 10%, which leads to a significant waste of information.[21]

At present, for computer processing, the most widely-used algorithm for SAR target detection is the Constant False Alarm Rate (CFAR),[22] which is based on the probability distribution model proposed by the American Lincoln laboratory. This method detects targets according to the difference in a statistical model between the target and the background clutter. Since the observation scene is different, the statistical model of clutter is also different. The commonly used statistical models include the Gaussian, Weibull, gamma lognormal, Rayleigh distribution model, etc. However, because the model and clutter are not fitted well, the detection accuracy of the CFAR is not high at the low SCR, and false alarms are more common.

Manual visual interpretation is believed to be the most reliable method for SAR image interpretation, but it is labor-intensive and time-consuming. In practice, the radar echo of the man-made target is higher than that of the background, and the features of intensity, orientation, texture and so on between them is also very different. Hence, it can easily catch the human visual system's attention spontaneously. On the other hand, the reliability of manual visual interpretation after professional training signifies the efficacy of the visual attention model in simulating goal-driven top-down cues in the human visual system, for target detection from SAR images. Hence, we try to apply the visual attention model to generate saliency maps to detect SAR image targets. In fact, the scattering characteristics of targets are determined by the wavelength of the electromagnetic wave, incident angle, polarization mode, surface structure, material of targets and so on. It is not only the basis of SAR target detection and classification, but also of SAR image interpretation.[23] A SAR image is an energy mapping of the backscattering characteristics of ground objects. It reflects their two-dimensional scattering characteristics, since it is obtained by a complex imaging algorithm after receiving the scattering echoes of ground objects. The scattering characteristics of the target determine its performance in SAR images, such as intensity characteristics, orientation characteristics and so on. However, there is a major difference between the characteristics of optical images and SAR images, hence many of the existing vision models based on optical images are not suitable for SAR images. First, the coherence between radar echo signals lead to strong speckle noises in the SAR images. Moreover, with only the grayscale channel, SAR images carry less information than optical images. The highest resolution of SAR images is about 0.1 meters, far lower than current optical images, and lower resolution means less effective features. To this end, we need to be very careful and make modifications if visual models are used to handle SAR images.

Tian et al.[24] improved the Itti model by combining multiple feature fusion and saliency computations, and applied the improved model to ship detection at sea surface. Peak and Yao[25] introduced the concept of group targets in SAR image target detection, specifically analyzing and using the characteristics of group targets in SAR images, which improved the detection rate and reduced the false alarm rate. By introducing the element based on neural biology and information entropy, Chen et al.[26] reduced some uncertain factors in the visual attention model, and achieved good results by optimizing the Itti model based on the features of targets in the SAR images.

Although the above methods improved traditional visual models according to the characteristics of SAR images, their scene adaptability is generally not strong. In different scenes, the results of experiments vary greatly under the same parameters, hence the models are not robust for automatic target detection or recognition for use in the SAR system. In addition, the above methods do not fully utilize the prior information of targets in images, and fail to fully grasp the essential features of the target highlighted from the background, resulting in the false alarm rate of the algorithm being too high under the low signal to clutter ratio. Therefore, for the visual model applied to target detection or recognition of SAR images, visual models for traditional optical images need to be modified according to the characteristics of targets and radar images, along with further exploitation of the human visual system in radar images interpretation. In particular, for top-down cues, visual signals are interpreted in a customized manner to serve the task demands. The top-down attention is usually task-driven, i.e., for different tasks top-down models should be specifically tuned. Usually, these models are used to detect or recognize targets in real-life visual scenes, i.e., faces, pedestrians, vehicles and so on. Therefore, we need to design a novel and targeted top-down model to adapt to the task and the knowledge of SAR images.

In this paper, a task-driven visual attention model is proposed to effectively and accurately detect vehicle targets in SAR images, which combines a top-down stage and a bottom-up stage. Our contributions lie in the following:

(1) Specifically, the bottom-up stage makes some simplifications and modifications to the traditional Itti model according to the characteristics of SAR images. We design a new weighting function facilitating the identification of multiple targets.
(2) During the top-down stage, we propose a novel goal-driven learning strategy to learn optimal parameters from the training set. These parameters can optimize the integration of various feature maps during the generation of the top-down saliency map, and maximise the saliency of targets. The top-down and bottom-up saliency maps are then integrated to acquire global saliency maps.
(3) In addition to being employed to obtain the best weights, the training dataset provides thresholds based on the average area size or lengths, thus serving as a priori knowledge of features. Finally, the detection result is obtained through the binarization and thresholding processes for the global saliency map.

The remainder of this paper is structured as follows. Section 2 presents the details of the proposed method, including a bottom-up and top-down process, modification of the Itti model for SAR target detection and a novel learning strategy. The design and results of experiments are detailed in Section 3, including efficiency and robustness comparisons. Finally, some concluding remarks and future research directions are outlined in Section 4.

2. Proposed method

In order to model the action mechanism of the human visual system, we propose a composite method which consists of two processes: the bottom-up and top-down process. The goal of the proposed method is to generate a saliency map with higher target saliency, so the computation process of this model algorithm is focused on generating a saliency map. Corresponding to the visual attention system cues in the human brain, the final feature map includes a bottom-up saliency map, a top-down saliency map, and ultimately, they are combined to form a global saliency map. In Fig. 1, we depict the complete framework of the proposed method. In this flowchart, the grey boxes denote our innovative work. The input image first passes the bottom-up process. In this process, we use the modified Itti model based on the SAR image characteristics, after which the Bottom-Up (BU) saliency map is obtained. In the top-down stage, by using the novel learning strategy designed for the task of SAR images target detection, optimal weights are learned. In this way, the sub saliency maps are weighted to integrate the top-down saliency map using the optimal weights. The generation progress of the sub saliency maps is the same as that in the bottom-up stage. Then, the single global saliency map is obtained by the linear integration of the BU saliency map and the Top-Down (TD) saliency map. Finally, the decision stage is executed, in which the statistical prior knowledge accumulated from the training set is used as thresholds to filter targets and exclude non target areas. A detailed description on the generation of these saliency maps is included in the following.

2.1. Bottom-up attention

Since the well-known Itti model has strong validity and credibility, the bottom-up process of the proposed method is based on it. In order to explain fully the whole process of the proposed method, a brief description of the Itti' specific process is given below.

2.1.1. Basics of Itti model

The processing of the Itti model[5] is divided into three steps: feature extraction, feature integration and saliency map generation. First, we construct a multi-scale map pyramid of brightness, color and direction according to the input image. Based on that, the feather maps of the three channels are calculated by the center-surround differential operation. Then, the sub saliency maps, indicating the significant intensity of each feature channel, are obtained by weighting the feature maps. Finally, they are normalized and combined to form a saliency map.

(1) Construct the image pyramid, and obtain the multi-scale maps of each feature channel

The original image s_0 is the bottom layer of Pyramid. First, Gauss low-pass filtering is performed on the original image,

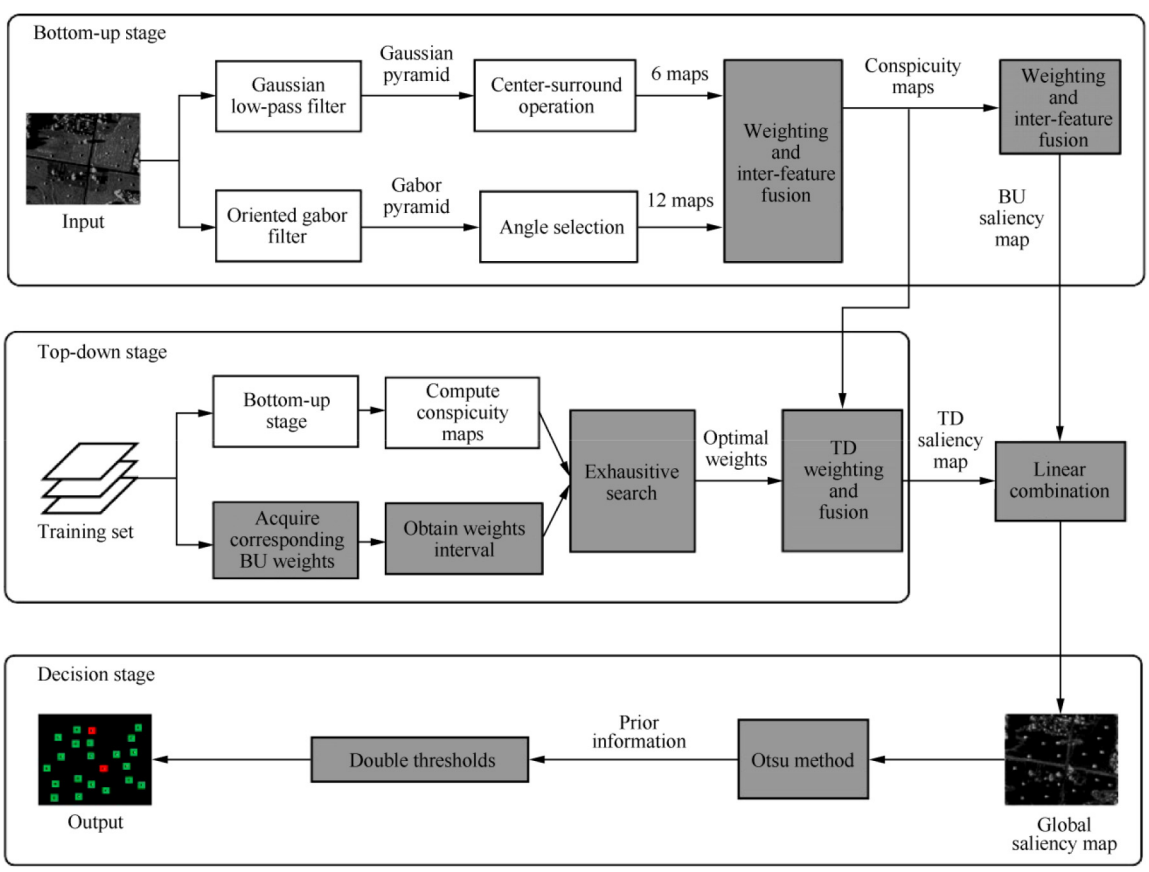


Fig. 1 Framework of proposed method.

and then the second level of Pyramid s_1 is obtained by 1/2 down-sampling. Repeat the above steps 8 times to get the 9 level of Gauss Pyramid s_σ ($\sigma \in \{0, 1, ..., 8\}$). In this process, it should be noted that filtering operations must be performed before sampling in order to smoothing the image and suppressing the noise.

The intensity, color and orientation scale map is calculated on the basis of Pyramid image.

Intensity scale map: intensity scale map of the σth level is:

$$I_\sigma(x, y) = \frac{(r + g + b)}{3} \tag{1}$$

Among them, (x, y) is any pixel in the image, and r, g and b are the red, green and blue components of this pixel respectively. By this process, 9 intensity scale maps are obtained.

Orientation scale map: Unlike the intensity and color channels, the image Pyramid of orientation feature is obtained by means of a directional Gabor filter. Using the Gabor filter with a direction of $\theta = \{0°, 45°, 90°, 135°\}$ to filter from the first level $I_{\sigma=0}(x, y)$ of intensity Pyramid, and 36 scale maps $O(\sigma, \theta)(\theta \in \{0°, 45°, 90°, 135°\})$ are obtained.

Color scale map: Because our subsequent method does not use color information, this part is omitted in this paper.

(2) Calculate the feature map of each feature channel

In the Itti model, the center-surround differential operation is defined to simulate the center-surround effect of receptive field in the retina. In the model, the scale map whose level is $c \in \{2, 3, 4\}$ is chosen as the center, and the scale map whose level is $s = c + \alpha$ ($\alpha \in \{3, 4\}$) is chosen as the surround. When calculating, the surround image is interpolated and enlarged to the same scale as the center image, and then point-to-point subtraction is carried out. The intensity feature map is defined as:

$$I(c, s) = |I(c) \Theta I(s)| \tag{2}$$

where $I(c)$ and $I(s)$ represent the intensiy scale map of cth and sth level, and the symbol Θ represents cross scale subtraction. 6 intensity feature maps are obtained after calculation. The calculation formula of the orientation feature map is as follows:

$$O(c, s, \theta) = |O(c, \theta) \Theta O(s, \theta)| \tag{3}$$

where $O(c, \theta)$ and $O(s, \theta)$ represent the orientation scale map of cth and sth level in the direction of θ. 6 feature maps can be obtained at each angle, so there are 24 orientation feature maps. We still omit the description of color maps.

(3) Integration features, and get significant map

If each feature map is integrated in a directly addition way, the contribution of the feature map to the saliency map is almost the same. Therefore, normalization function $N(\cdot)$ is used to normalize the feature maps in the Itti model, so that the feature map with a few strong points has greater weight. The normalization function is defined as:

$$N(X) = \left(M - \bar{m}\right)^2 X \tag{4}$$

When calculation, we first normalize the feature maps to the fixed interval $[0, M]$, and $\bar{m}$ is the mean of the local maximum. Therefore, the sub saliency map corresponding to each feature channel is:

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c, s)) \tag{5}$$

$$\bar{O} = \sum_{\theta=\{0°,45°,90°,135°\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c, s, \theta))\right) \tag{6}$$

The final feature map is obtained by further normalization and linear superposition of three sub feature maps:

$$S = \frac{1}{3}\left(N\left(\bar{I}\right) + N\left(\bar{C}\right) + N\left(\bar{O}\right)\right) \tag{7}$$

where $\bar{C}$ is sub saliency of color channel.

2.1.2. Modification of Itti model for SAR target detection

The Itti model was not used directly, as this would result in waste of computation and problems with adaptability. Therefore, we made some modifications according to the characteristics of SAR images, which have unique imaging principles. SAR images have only one channel, compared to the three channels of RGB like optical images, and no color information in their images, hence we ignore the color channel in the Itti model. Intensity and orientation channels are mostly consistent with those in the original Itti model, but they possess some specific modifications based on the characteristics of the targets in SAR images. The flowchart shows that the bottom-up process mainly consists of three stages, i.e., the feature extraction stage, the weighting operation stage and the saliency map generating stage. The three parts of the bottom-up process is described in detail below.

(1) Feature Extraction
(A) Intensity channel

For an input image I, we first use the Gaussian filter and sub-sampling on its intensity to generate a five-scaled Gauss Pyramid. Then, after a center-surround operation, the Gaussian image pyramid with five scales s_0-s_4 is further transformed into the feature maps. There are 9 scales in the image pyramid structure of the Itti model, but there are only 5 scales in that of our method. There are two main reasons for this choice. One is less computation and time consumption, and the other is that the resolution of a saliency map is usually inversely proportional to the level of scale. For the SAR images with vehicle targets, the vehicle possesses only hundreds of pixels, and higher scales lead to lower resolution. In what follows the steps for intensity feature extraction are detailed.

Step 1. Generate five-scaled Gaussian pyramid I_σ, where $\sigma \in \{0, 1, 2, 3, 4\}$ represents the scale. The low-pass Gaussian filter is expressed as the following formula:

$$G(x, y, o) = \frac{1}{2\pi o^2} \exp\left(-\frac{x^2 + y^2}{2o^2}\right) \tag{8}$$

where (x, y) denotes the coordinate of an arbitrary pixel and O is the standard deviation which we set $o = 3$ here.

Step 2. Represent surround of arbitrary pixel.

For each pixel $I''_{\sigma=c}(x, y)$ in the center, its surround is represented as:

$$I''_{\sigma=c|\delta}(x,y)=\left[\sum\left(I_{s_i}(x-\delta,y-\delta)+I_{s_i}(x-\delta+1,y-\delta)+I_{s_i}(x-\delta,y-\delta+1)+\cdots+I_{s_i}(x+\delta,y+\delta)\right)\right]/\delta^2 \tag{9}$$

where $c\in\{2,3,4\}$ is the center, $\delta\in\{4,8\}$ is the edge length of its rectangular neighborhood.

Step 3. Preform the center-surround operation and generate 6 feature maps $I''_{c,\delta}$.

The feature maps are defined as:

$$I''_{c,\delta}=I''_{\sigma=c}-I''_{\sigma=c|\delta} \tag{10}$$

Through the above steps, 6 intensity feature maps are finally acquired.

(B) Orientation channel

Although the targets appear very small in many SAR images, the orientation information is still indispensable when detecting targets, particularly since the resolution of SAR images is high.[27,28] In the Itti model, the orientation feature is obtained from the orientated eight-scaled Gabor pyramid. As discussed previously, the targets in SAR images possess only hundreds of pixels and less so for higher scaled images. Therefore, we accept the operation in Eq. (8) where only 5 scales are needed. When the orientation feature is extracted, the center-surround operation is discarded, as the oriented center-surround difference has been applied implicitly in the Gabor filter. At the same time, this approach can also prevent ambiguity of the feather map caused by the center-surround operation. The oriented Gabor pyramid is defined as:

$$H(x,y,\sigma,\theta)=\frac{1}{\sigma^2}\exp\left(-\pi\frac{x^2+y^2}{\sigma^2}\right)\times\left\{\exp[\mathrm{i}2\pi(x\cos\theta+y\sin\theta)]-\exp\left(-\frac{\pi^2}{2}\right)\right\} \tag{11}$$

where $\sigma\in\{2,3,4\}$ is the scale and $\theta\in\{0^\circ,45^\circ,90^\circ,135^\circ\}$ is the preferred angle. Therefore, the orientation feature map $O''_{\sigma,\theta}(x,y)$ can be obtained by applying Eq. (11). Finally, through these steps, 12 original orientation feature maps are generated.

(2) Normalization and weighting

After obtaining the intensity and orientation feature maps, they are normalized to the same scale, and then fused respectively to form two sub saliency maps. Different features have different contributions to the perceptual saliency, we have to determine the most important maps and raise their influence. This can be achieved by an operator like the normalization operator presented in the Itti model, $N(X)=X(M-\bar{m})$, where X represents the feature map, M is the global maxima, and $\bar{m}$ is the expectation of local maxima. However, there is a lethal problem with this approach in identifying targets. For instance, if it is aimed at detecting several vehicles within an image, a feature map may have several nearly equal maxima which indicates the targets, but the normalization operator will only yield zero or a very small value, i.e., the feature map contributes almost nothing. This is of course unacceptable. Therefore, in order to prevent these problems, a new weighting function is proposed to ensure that multiple targets are highlighted simultaneously. It has been tested and proved to be effective and credible, hence we replace the $N(X)$ in the Itti model with it. The weighting function is defined as:

$$W(X)=\frac{1}{\sqrt{2\pi}\eta}\mathrm{e}^{\frac{-(M-\bar{m})^2}{2\eta^2}}\cdot\frac{\bar{m}}{\bar{r}}\cdot X \tag{12}$$

where X, M, and $\bar{m}$ are defined above. η is the standard deviation of the feature map, and $\bar{r}$ is the expectation of the rest of the feature map when taking out the local maxima.

So the conspicuity map for each channel is achieved by first weighting the corresponding feature maps and then applying across-scale addition:

$$I'=\oplus_{c,\delta}W\left(I''_{c,\delta}\right) \tag{13}$$

$$O'=\oplus_{\sigma,\theta}W\left(O''_{\sigma,\theta}\right) \tag{14}$$

where $\oplus$ denotes point-to-point addition.

(3) Bottom-up saliency map

The sub saliency map obtained from each feature channel also needs to be weighted to form the bottom-up saliency map. The saliency map indicates the regional saliency of an image, and the highly significant target or region presents higher brightness in the image. The generation process of the saliency map is represented as:

$$S_{\mathrm{BU}}=W(I')+W(O') \tag{15}$$

2.2. Top-down attention

Unlike the bottom-up cues which are mainly affected by the appearance characteristics of a visual scene (spontaneous stimulus-driven), top-down cues are determined by the cognitive phenomena of the human brain, such as knowledge, expectation, reward, current goals, etc. Before attaining professional training, it is very difficult for humans to distinguish whether a region in a SAR image is a target (such as a vehicle) or not. In terms of visual perception, vehicles in SAR images are extremely different to those in optical images. In addition, the usual SAR images have a lower resolution than optical images, which makes the task more difficult. However, if observers are given the opportunity to observe SAR target images and receive specialist training in advance, it will then be easier for them to recognize targets in a SAR scene. In addition to the targets' self-existent low-level characteristics, such as intensity and orientation, which makes them more significant and attracts attention, the tasks and goals also have a guidance impact on the visual attention search process.

Furthermore, prior knowledge information such as area, outline and texture play important roles in the human's visual understanding process for images. Therefore, the top-down model should be constructed and designed specially by the particular task. Appropriate use of prior information is very beneficial and even crucial to accurately detect vehicle targets.

The top-down process in the visual model proposed in this paper belongs to the first class in the top-down model classification explored in the introduction, i.e. it is a process of the

weighted integration for the low-level visual features using the cognitive phenomenon in the top-down manner. In this paper, the top-down stage is also based on the saliency map generated by the optimal combination of weights acquired by the novel learning strategy, and prior knowledge such as length and area size of targets helps to make further decisions.

2.2.1. Learning strategy

In the final stage of the bottom-up process, a saliency map is generated from two sub saliency maps based on Eq. (15). Nevertheless, the weights computed from Eq. (15) are not necessarily optimal for meeting the goal of targets being most salient in saliency maps. Therefore, our goal is to generate the most appropriate saliency map, i.e. to find the weights that are more advantageous to target detection than the former ones. The learning strategy is designed to obtain optimal weights as far as possible. We propose the following learning strategy whilst considering the task of target detection, which includes an exhaustive search and quantitative evaluation.

The general process of the learning strategy is depicted in the middle of Fig. 1. In order to complete the learning process, we first need a set of image slices containing targets as the training set. For slice X_i, (the subscript i denotes the order of slices), two corresponding conspicuity maps I_i and O_i are computed with the aforementioned bottom-up process. Instead of using the weighting function $W(\cdot)$ to form the saliency map, we generate different saliency maps of each image slice by exhaustive searching for different weights. Then we use F-measure to evaluate quantitatively every map and choose the most accurate one we need to adapt the task of target detection, and the weights of this map will be recorded. Below is the detailed steps.

Step 1. Reconsider Eq. (12), and let $w(X)$ denote the item $\frac{1}{\sqrt{2\pi}\eta}\mathrm{e}^{\frac{-(M-\bar{m})^2}{2\eta^2}}\cdot\frac{\bar{m}}{\bar{r}}$, i.e.:

$$w(X)=\frac{1}{\sqrt{2\pi}\eta}\mathrm{e}^{\frac{-(M-\bar{m})^2}{2\eta^2}}\cdot\frac{\bar{m}}{\bar{r}} \tag{16}$$

For each slice X_i, compute the bottom-up weights of the conspicuity maps: $w_{I_i}=w(I_i)$, $w_{O_i}=w(O_i)$. Assume there are N training slices in total, and therefore we get two sets of weights $\{w_{I_1},w_{I_2},\cdots,w_{I_N}\}$ and $\{w_{O_1},w_{O_2},\cdots,w_{O_N}\}$ corresponding to the intensity conspicuity maps and orientation conspicuity maps respectively.

Step 2. Determine the intervals of top-down weights $[w_{I_{\min}},w_{I_{\max}}]$ and $[w_{O_{\min}},w_{O_{\max}}]$. The interval is defined as:

$$w_{I_{\min}}=\min(w_{I_i})-\sigma_I \tag{17}$$

$$w_{I_{\max}}=\max(w_{I_i})+\sigma_I \tag{18}$$

$$w_{O_{\min}}=\min(w_{O_i})-\sigma_O \tag{19}$$

$$w_{O_{\max}}=\max(w_{O_i})+\sigma_O \tag{20}$$

where σ_I and σ_O are the standard deviation of w_{I_i} and w_{O_i}, respectively.

Step 3. Select 10 weights from every interval at a regular distance and thus 100 saliency maps can be generated by the 100 weight groups for each target slice.

Step 4. Benchmark the 100 saliency maps and find the best one with its corresponding weights w'_{I_i} and w'_{O_i}. Here we use the Precision (P), Recall (R) and F-measure (F) as the benchmarks, given by

$$P=\sum(S\otimes G)\Big/\sum(S) \tag{21}$$

$$R=\sum(S\otimes G)\Big/\sum(G) \tag{22}$$

$$F=(\alpha^2+1)PR/[\alpha^2(P+R)] \tag{23}$$

where S is the saliency map, G is the ground-truth segmented manually. The operator $\otimes$ means point-to-point multiplication.

Step 5. Eventually we get the best pair of (w'_{I_i},w'_{O_i}) for each target slice, and the means of the two sets of weights, coming from N target slices, are thought to be the desired weights of the targets:

$$w_I=\sum_{i=1}^{N}w'_{I_i}/N \tag{24}$$

$$w_O=\sum_{i=1}^{N}w'_{O_i}/N \tag{25}$$

In this way, the 100 different optimized combinations are combined together.

2.2.2. Top-down saliency map

After learning the optimal weight, the top-down saliency map is synthesized based on two bottom-up sub saliency maps and top-down weights.

$$S_{\mathrm{TD}}=w_I I'+w_O O' \tag{26}$$

It is worth noting that we only calculate the optimal weights for combining the two sub saliency maps together. In fact, this method can also be used to calculate the optimal weights for the raw feature maps, but they have an almost negligible effect in comparison with the resulting huge computational costs.

2.3. Global saliency map

The global saliency map is then generated from the combination of the bottom-up and top-down maps. Parameter t_{TD} determines how much the top-down process contribute to the global saliency map. For real application, we choose an empirical value of 0.5 for the parameter t_{TD}. This empirical value balances the two processes, so that the proposed method can produce effective detection performance in most practical scenarios and under various signal to clutter ratio conditions. If t_{TD} is set too large or too small, the salient map will be worse, leading to degradation of target detection performance. The precise selection of t_{TD} involves the fusion process of top-down and bottom-up stages, which is known to be a highly complex process in the human visual system and has yet to be fully studied. At present, we can still use the learning method outlined above to obtain the optimal weight. However, in case the empirical value is very effective, a huge computational cost will be introduced. Finally, we use the empirically estimated value t_{TD} as an alternative.

$$S=(1-t_{\mathrm{TD}})S_{\mathrm{BU}}+t_{\mathrm{TD}}S_{\mathrm{TD}} \tag{27}$$

2.4. Decision-making process

In the top-down process, the optimal weights learned from target slices generating the TD salience map, is the core of this part of the methods. In addition, the area and length of a specific type of vehicle as prior knowledge information also play a crucial role in target detection tasks. In the final decision stage, we choose the area represented by the number of pixels a vehicle possesses and the length of it as two thresholds to filter targets and exclude non-target areas. First, we transform the saliency map into a binary map. Here we use the Otsu method[29] to create the global threshold T to extract the salient regions. The Otsu method is an adaptive threshold segmentation method. The threshold T is selected in the gray range, and the image is divided into foreground and background to be binarized. Then the between-clusters variance is calculated. Searching for T until T is found to maximize the between-clusters variance, and at this point T is the optimal threshold.

$$T = \mathrm{Arg}\max_{0\le t\le L-1}[p_1(t)(\mu_1(t)-\mu)^2 + p_2(t)(\mu_2(t)-\mu)^2] \tag{28}$$

where L is the gray level of an image, and $p_1(t)$ and $p_2(t)$ are the ratio of foreground and background pixels to total image pixels when the threshold is t. $\mu_1(t)$ and $\mu_2(t)$ are the pixel mean value of foreground and background, and μ is the pixel mean value of the whole image. The formula for binarization is as follows:

$$S_{\mathrm{bw}}(x,y) = \begin{cases} 1 & S(x,y) > T \\ 0 & S(x,y) \leqslant T \end{cases} \tag{29}$$

For an arbitrary region in the binary map, it is determined whether it is a target or not by the area A and length L of the target:

$$R_i = \begin{cases} 1 & A \in [a,b] \text{ and } L \in [c,d] \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

where R_i is the suspicious regions, 1 for target, 0 for not. The confidence intervals $[a,b]$ and $[c,d]$ are computed from the ground-truths of training slices in the learning stage.

3. Experimental results and analysis

The SAR images used in experiments are from the Moving and Stationary Target Acquisition and Recognition (MSTAR) database. This benchmark data was collected using the Sandia National Laboratories Twin Otter SAR sensor payload, operating at X-band with a high resolution of 0.3 m, spotlight mode, and HH single 320 polarization. We picked 80 cluttered images from spotlight SAR images in the MSTAR database with an image size of 1478×1784 pixels. Each image was added with 20 targets from classes of type BRT70, T72 and BMP2. As the testing set, the 20 target slices in the 80 scene images were different from the N target slices used as the training set.

In comparison, six state-of-the-art methods were applied to the SAR target sets. One method is the CFAR[30] which is a well-known method for SAR image detection. The other three saliency detection methods are IG,[31] the Spectral Residual (SR)[32] and SIM,[33] which are also efficient visual models different from Itti's. The IG method achieves better image integrity by frequency tuning. Firstly, the image is transformed into Lab color space and filtered by Gauss filter. Then, the mean of each channel and the Euclidean distance between them is calculated separately. By this way, the saliency map of the image is obtained. The SR method analyzes the visual saliency of images in the respect of frequency spectrum. The SIM obtains a saliency model by using a principled selection of parameters as well as an innate spatial pooling mechanism. To ensure fairness and the reasonableness of the experiment, not only in CAFR, but also in other contrast visual model methods, we used the same decision-making process as what is used in the proposed method.

Furthermore, Object Proposal (OP)[34] is adopted. This aims to cover as many objects of interest as possible with as fewest windows, by segmenting images into superpixels firstly, and then specific strategies are used to aggregate superpixels into objects. In addition, the YOLO_v2,[35] a recent and efficient detection algorithm in Deep Learning, is also used as a comparison method. The YOLO_v2 is currently the state-of-the-art method for standard detection tasks like PASCAL, VOC and COCO. The YOLO_v2 forms the object detection as a regression problem to the spatially separated bounding boxes and associated class probabilities. The YOLO_v2 is the supervised algorithm, and its straining has two steps. First, some ground truth files (e.g. classes, positions, shapes of targets in each SAR image) are generated; then, these files are used to finely tune the YOLO_v2 pre-trained on the Imagenet dataset. Considering the network structure of the YOLO_v2, we randomly embed the vehicle targets into 600×600 scenes as the training set, in which data augmentation strategies such as shift and rotation are adopted to extend the training samples. The ground-truth of training chips is manually labelled. It is worth noting that the corresponding shadow regions and some extra background areas are included, which is extremely important. Since Deep Learning algorithms have large requirements for the data quantity of the labeled samples, in order to train the YOLO_v2 more adequately, we made 3000 scene images as the training set. Because of the difference of algorithm mechanisms and the training method between the YOLO_v2 and the proposed method, the training set of the YOLO_v2 is scene images rather than target slices and the amount of samples in the training set is a lot more than that of the proposed method. Moreover, we note that the proposed method has better performance than the method utilizing the bottom-up process only, which demonstrates the effectiveness of the top-down learning strategy.

To demonstrate the robustness and effectiveness of the proposed method, we also performed experiments under different Signal to Clutter Ratio (SCR) conditions.

3.1. Benchmarks

There are several benchmarks to quantitatively evaluate the effectiveness of a detection algorithm. Among them the probability of detection (P_{d}) and probability of false alarm (P_{f}) are the most frequently used, hence we choose them to evaluate our method and to facilitate comparison with others. P_{d} and P_{f} are defined as:

$$P_{\mathrm{d}} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{31}$$

$$P_{\mathrm{f}} = \mathrm{FP}/(\mathrm{TP}+\mathrm{FP}) \tag{32}$$

where TP means the number of detected targets, FN denotes the number of true targets that are missed, and FP is the number of false alarms.

Besides, the Precision, Recall and F-measure are the fundamental measures in statistics, therefore they are also included in our experiments. In our cases, Recall (R) has the same definition as P_{d}. Precision (P) and F-measure (F) are defined as:

$$P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{33}$$

$$F = \left(\alpha^2 + 1\right)PR/\left[\alpha^2(P + R)\right] \tag{34}$$

As can be seen in Eq. (34), the F-measure is the weighted harmonic mean of Precision and Recall.[36] In our experiments, it was set as $\alpha = 1$. In this experiment, the final results are the average or steady results after the repeated experiments.

Because the proposed method uses learning strategy, which belongs to supervised algorithm, we need to prepare training set in advance. We selected 100 vehicle target slices of BRT70 from the MSTAR database as the training set. The training set is also used to determine the confidence interval mentioned in Eq. (30), calculated as [35.15, 46.40] and [420.30, 484.34] using Eq. (35), with the confidence probability of 93.5% and 92.3%, respectively. The interval $[a, b]$ and $[c, d]$ are defined as:

$$\begin{aligned} a &= \overline{\mu_s} - \sigma_s,\ b = \overline{\mu_s} + \sigma_s \\ c &= \overline{\mu_l} - \sigma_l,\ d = \overline{\mu_l} + \sigma_l \end{aligned} \tag{35}$$

where $\overline{\mu_s}$ and $\overline{\mu_l}$ are the expectations of the area and length of each training target, σ_s and σ_l are the relevant standard deviations.

3.2. Two representative scenes

In order to illustrate the comparison and results of experiments, we selected two representative scenes from the 80 experimental scenes shown in the paper. They are shown in Figs. 2 and 3. The former is a scene with slight distracters and the latter is a scene with heavy distracters.

3.2.1. Scene with slight distracters

It is noted from Fig. 2 that the vehicles in this image are distinct from the surrounding and thus possess strong conspicuity. The Bottom-Up (BU), Top-Down (TD) and global saliency maps are shown in Fig. 4. Evidently, the targets in the top-down saliency map are more identifiable from their surroundings than they are in the bottom-up saliency map, which validates the effectiveness of our top-down approach.

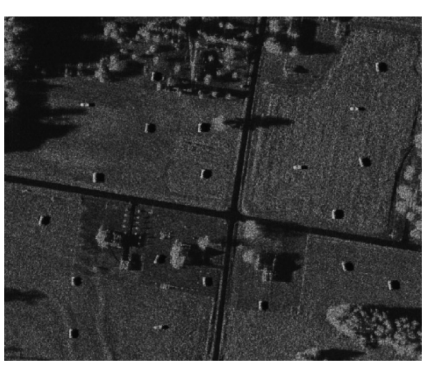


Fig. 2 Scene 1 with 20 vehicle targets inside.

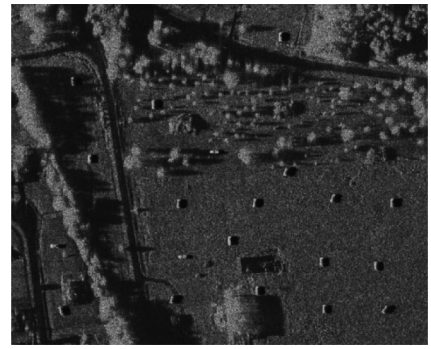


Fig. 3 Scene 2 with 20 vehicle targets inside.

Fig. 5 illustrates detection results obtained in this experiment. The green rectangles mark the detected targets, while the red and white ones mark the false alarms and undetected targets, respectively. The principles of the OP and YOLO_v2 are different from other methods. There is no binarization process, so its results are in different form. The IG method detected 15 targets and missed 5 targets, and no false alarm appeared. The SR detected 14 targets and also did not generate any false alarm. The SIM detected 16 targets and introduced 6 false alarms. In contrast, our method detected 19 targets and generated only 1 false alarms. To evaluate the proposed method in depth, the result of the BU only was presented, which missed 4 targets and generated 1 extra false alarm. The CFAR, the most commonly used method for SAR image target detection, detected 16 targets and generated 2 false alarms, and the YOLO_v2 detected 19 targets but introduced 6 false alarms. The performance of OP is the worse, and it just detected 9 targets.

A quantitative comparison of the experimental results is shown in Table 1. As can be seen, except for the false alarms of the proposed method being slightly higher than the SR's, the other indicators are all consistently best. Comparably, the BU, CFAR, IG, SR, SIM, OP and the YOLO_v2 lag behind our method by 8.51%, 9.29%, 10.79%, 9.29%, 18.81%, 32.93% and 10.56% when using $F_{\alpha=1}$ metrics.

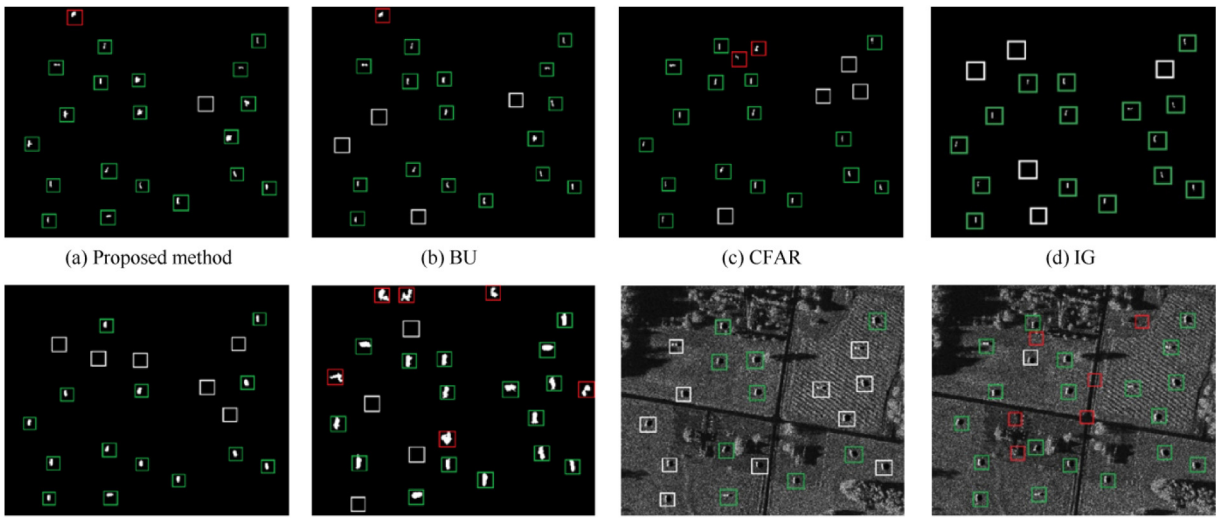
3.2.2. Scene with heavy distracters

In this scene, the image is more cluttered than the former by possessing less flat regions and more distracters. The saliency maps and detection results are shown in Figs. 6 and 7, respectively. The proposed method achieved the best detection result though the experimental result is not as good as the result in scene 1. Specifically, there were two targets undetected and one false alarm for the proposed method, whereas the UB approach missed four targets and generated 1 false alarm. The CFAR had four targets undetected and generated one false alarm. IG produced better performance and detected 17 targets with no false alarm. SR had five targets undetected and did not generate false alarms, and the SIM missed seven targets and introduced 16 false alarms. OP detected 12 targets and generated one false alarm, while the YOLO_v2 missed one target and had four false alarms.

Table 2 shows quantitative evaluation for the two methods. The BU, CFAR, IG, SR, SIM, OP and the YOLO_v2 lag behind our method by 5.82%, 0.42%, 9.38%, 6.59%, 3.92%,

(a) BU (b) TD (c) Global

Fig. 4 Saliency maps of scene 1.



(a) Proposed method (b) BU (c) CFAR (d) IG

(e) SR (f) SIM (g) OP (h) YOLO_v2

Note: Green rectangles mark detected targets; red and white ones mark false alarms and undetected targets respectively.

Fig. 5 Detection results of the contrast methods and the proposed method of scene 1.

Table 1 Quantitative measurement obtained by contrast methods and proposed method for scene1.

Method	P_d (%)	P_f (%)	P (%)	R (%)	$F_{\alpha=1}$ (%)
Proposed method	95	5	95	95	95
BU	80	5.88	94.12	80	86.49
CFAR	80	11.11	80.89	80	84.21
IG	75	0	100	75	85.71
SR	70	0	100	75	85.71
SIM	80	27.27	72.73	80	76.19
OP	45	0	100	45	62.07
YOLO_v2	95	24	76	95	84.44

19.27% and 3.94% using $F_{\alpha=1}$ metrics. In addition, what merits our attention is that YOLO_v2 also has a powerful target detection capability, but generates more false alarms. Finally, IG also performed well.

3.3. Robustness analysis

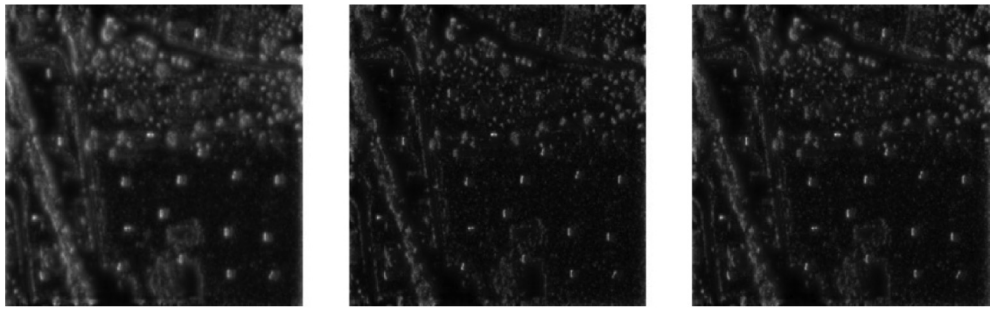
To further analyze the robustness of the above methods, we carried out the experiments under different SCR conditions by tuning the instance of targets, which is calculated as follows:

$$\mathrm{SCR} = 20\lg\frac{\bar{I}_{\mathrm{tar}}}{\bar{I}_{\mathrm{bcg}}} \tag{36}$$

where $\bar{I}_{\mathrm{tar}}$ and $\bar{I}_{\mathrm{bcg}}$ denote the mean intensities of the target and its surrounding background. We still adapt the $F_{\alpha=1}$ to execute the comprehensive evaluation. Under each different SCR condition, we carried out experiments in 80 scenes and computed the average $F_{\alpha=1}$ of every scene for the embodiment of algorithm detection performance. After a number of experiments, we obtain a curve of $F_{\alpha=1}$ over SCR, as shown in Fig. 8.

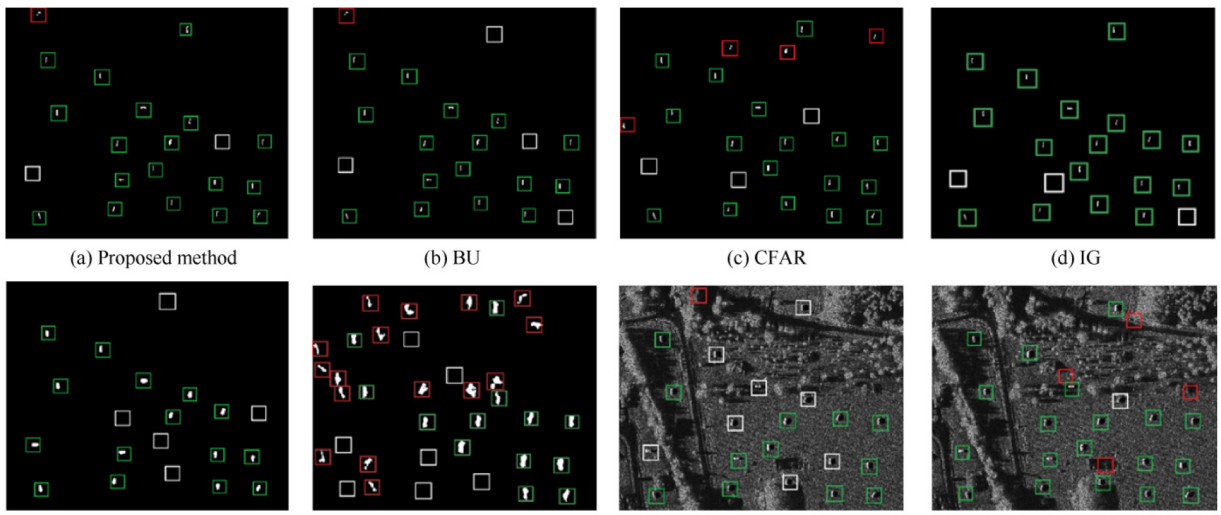
In order to analyze the distribution of experiment samples under the different SCR conditions from a statistical point of view, we calculated the mean and variance of $F_{\alpha=1}$ in the robustness experiments. The statistical results are shown in Table 3.

We find that the target detection performance of the proposed method is the best when the SCR is large. From Table 3, we can see that the average detection performance of the proposed method is superior to the other comparison experiments under the different SCR conditions. The variance of the proposed method is only inferior to YOLO_v2, which is consistent



(a) BU (b) TD (c) Global

Fig. 6 Saliency maps of scene 2.



(a) Proposed method (b) BU (c) CFAR (d) IG

(e) SR (f) SIM (g) OP (h) YOLO_v2

Note: Green rectangles mark detected targets; red and white ones mark false alarms and undetected targets respectively.

Fig. 7 Detection results of contrast methods and proposed method of scene 2.

Table 2 Quantitative measurement obtained by contrast methods and proposed method for scene 2.

Method	P_d (%)	P_f (%)	P (%)	R (%)	$F_{\alpha=1}$ (%)
Proposed method	90	5.26	94.74	90	92.31
BU	80	5.88	94.12	80	86.49
CFAR	85	19.05	80.89	85	82.93
IG	85	0	100	85	91.89
SR	75	0	100	75	85.71
SIM	65	55.17	44.83	65	53.06
OP	60	7.69	92.31	60	72.73
YOLO_v2	95	17.39	82.61	95	88.37

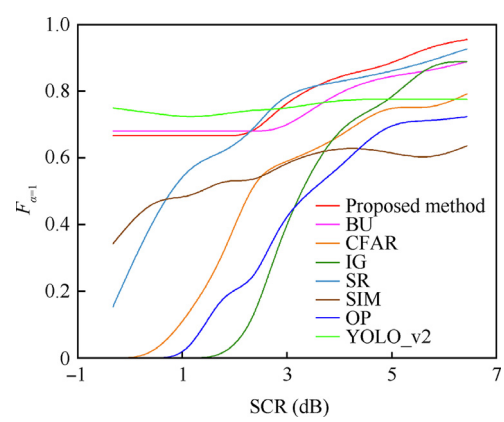


Fig. 8 $F_{\alpha=1}$ of eight methods under different SCR.

with the BU way. When the SCR is reduced, the performance of the proposed method is still more stable than the CFAR, IG, SR, OP and SIM, roughly the same as the BU, so it has good robustness.

Although the detection performance of IG under high SCR condition is excellent and almost no false alarm occurs, when the condition of SCR deteriorates, its performance decreases sharply. In these experiments, the IG has the greatest variance, so its stability is the worst. The YOLO_v2 is the most robust because it belongs to Deep Learning algorithms and it can automatically extract multiple features. The reduction of SCR only affects the intensity feature, and it still extracts the features with little influence on the intensity. Our method extracts the intensity and orientation features manually, so

Table 3 Statistical results of $F_{\alpha=1}$ under different SCR conditions.

Method	$F_{\alpha=1}$	
	Mean (%)	Variance
Proposed method	81.08	0.0416
BU	78.48	0.0216
CFAR	39.58	0.3134
IG	44.44	0.3951
SR	54.06	0.2992
SIM	48.96	0.0431
OP	36.19	0.1620
YOLO_v2	76.27	0.0003

the decrease of target intensity will have an impact on the detection performance.

3.4. Comparison analysis

For the more general condition, when the SCR is large (for example equal to 6.4), which is shown in the last column in Fig. 8, we show the $F_{\alpha=1}$ result in Table 4 specially. As can be seen in Table 4, the BU, CFAR, IG, SR, SIM, OP and the YOLO_v2 lag behind our method by 6.11%, 15.83%, 6.11%, 2.26%, 31.36%, 22.61% and 17.45% using $F_{\alpha=1}$ metrics. Compared with IG, SR and SIM, our method produces enhanced performance. This shows that the proposed method is more suitable for SAR images target detection than general vision models, because of its ability to optimize characteristics of SAR images. Further, the proposed method with top-down cues improved detection performance when compared to the pure BU, and this is evidence of the fact that humans have greater capability to understand SAR scenes once trained. This is why our method is based on the human visual system and the use of cognitive phenomenon is deemed effective. The results of CFAR are not as good as the proposed method in this paper, which shows that the proposed method has great advantages in SAR image target detection, even compared with the most widely used SAR image target detection algorithm. The poor performance of OP in SAR image target detection may be related to reducing feature information contained in SAR images, which needs further verification. Because of the lack of sufficient training and pertinent optimization for SAR images, the YOLO_v2 method, whilst demonstrating excellent performance in detecting optical images, does not acquire the same satisfactory performance.

Table 4 Quantitative measurement under condition of SCR = 6.4.

Method	$F_{\alpha=1}$ (%)
Proposed method	95.00
BU	88.89
CFAR	79.17
IG	88.89
SR	92.74
SIM	63.64
OP	72.39
YOLO_v2	77.55

Furthermore, as we found in scenes 1 and 2, the YOLO_v2 is very sensitive to the scene. This may be due to the limited number of training samples that cannot cover all the background features. In addition, we also find that the YOLO_v2 cannot identify targets that are close to each other. The reason is that the input data will be first cut into a lot of grids when the YOLO_v2 detects targets, and each grid is assumed having 1 target only.

4. Conclusions

In this paper, a novel SAR image target detection method based on a visual attention model is proposed, combining bottom-up and top-down processes. At the bottom-up stage, the proposed method makes a specific modification to the traditional Itti model, according to the characteristics of SAR images and target detection tasks. At the top-down stage, we propose a novel learning strategy to learn the optimal weights needed to generate a saliency map. This learning strategy has outstanding generalization ability, as once the optimal weights obtained by the finite training set are learned, they play a very effective role in most scenarios. The novelty of our proposed method lies in the following three aspects. Firstly, the novel weighting function makes multiple targets more identifiable. Secondly, top-down cues are introduced to learn optimal weights from the training set. Finally, the target's prior information such as area and length are used as thresholds in the decision stage, which is reliable for a final decision. Simulation experiment results illustrate that the proposed method possesses greater capability and robustness in detecting vehicle targets in comparison with other state-of-the-art visual models and detection methods, such as CFAR and YOLO_v2. Furthermore, the new method performed better than a bottom-up only approach, which further validated the effectiveness of our top-down learning strategy.

Finally, whilst our proposed method is primarily designed to detect vehicle targets from SAR images, it can also be applied to other fields, for example, detection of other types of fixed targets and optical images. Future work will thus explore the potential of this cognitively-inspired method as a benchmark resource for the SAR research community, as well as its adaptability to other challenging target detection tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 61771027, 61071139, 61471019, 61671035). Dr. Yang is supported in part under the Royal Society of Edinburgh-National Natural Science Foundation of China (RSE-NNSFC) Joint Project (2017–2019) (No. 6161101383) with China University of Petroleum (Huadong). Professor Hussain was partially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) (Nos. EP/I009310/1, EP/M026981/1).

References

1. Frintrop S, Backer G, Rome E. Goal-directed search with a topdown modulated computational attention system. *27th DAGM symposium pattern recognition*; 2005 Aug 31–Sep 2; Vienna, Austria. Berlin, Heidelberg: Springer-Verlag; 2005.

2. Li Z. A saliency map in primary visual cortex. *Trends Cogn Sci* 2002;**6**:9–16.
3. Treisman A. Search, similarity, and integration of features between and within dimensions. *J Exp Psychol Hum Percept Perform* 1991;**17**(3):652–76.
4. Koch C, Ullman S. *Selecting one among the many: A simple network implementing shifts in selective visual attention*. Cambridge, USA: Massachusetts Institute of Technology; 1984, Report No.: AIM-770.
5. Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res* 2000;**40**(10–12):1489–506.
6. Goferman S, Zelnik-Manor L, Tal A. Context-aware saliency detection. *IEEE Trans Pattern Anal Mach Intell* 2012;**34**(10):1915–26.
7. Jiang CY. Top-down attentional guidance based on implicit learning of visual covariation. *Psychol Sci* 1999;**10**(4):360–5.
8. Goldberg ME, Bushnell MC. Behavioral enhancement of visual responses in monkey cerebral cortex. II. Modulation in frontal eye fields specifically related to saccades. *J Neurophysiol* 1981;**46**(4):773.
9. Wolfe JM. Guided search 2.0: A revised model of visual search. *Psychon Bull Rev* 1994;**1**(2):202–38.
10. Wolfe JM. Guided search 4.0: Current progress with a model of visual search. *Integr Models Cogn Syst* 2012;**1**(2):202–38.
11. Itti L. Feature combination strategies for saliency-based visual attention systems. *J Electron Imaging* 2001;**10**(1):161–9.
12. Navalpakkam V, Itti L. Search goal tunes visual features optimally. *Neuron* 2007;**53**(4):605–17.
13. Kouchaki Z, Nasrabadi AM, Maghooli K. A new approach of feature combination for object detection in saliency-based visual attention. *Int J Comput Appl* 2013;**61**(19):7–12.
14. Han B, Xinbo G, Walsh V, Tcheang LA. A saliency map method with cortex-like mechanisms and sparse representation. *ACM international conference on image & video retrieval*; 2010 Jul 5–7; Xi'an: China. New York: ACM; 2010.
15. Tsotsos JK. Toward a computational model of visual attention. *Early Vision & Beyond* 1995;**1**:207.
16. Kim B, Ban SW, Lee M. Growing fuzzy topology adaptive resonance theory models with a push-pull learning algorithm. *Neurocomputing* 2011;**74**(4):646–55.
17. Borji A, Ahmadabadi MN, Araabi BN, Hamidi M. Online learning of task-driven object-based visual attention control. *Image Vis Comput* 2010;**28**(7):1130–45.
18. Ban SW, Szu HH, Agee FJ, Lee M. Autonomous mental development with selective attention, object perception, and knowledge representation. *SPIE Defense and Security Symposium*; 2008 Mar 16; Orlando, USA. Bellingham: SPIE; 2008.
19. Yang J, Yang MH. Top-down visual saliency via joint CRF and dictionary learning. *IEEE conference on computer vision and pattern recognition*; 2012 Jun 16–21; Providence, USA. Piscataway: IEEE Press; 2012.
20. Zhang F, Yao X, Tang H, Yin Q, Hu Y, Lei B. Multiple mode SAR raw data simulation and parallel acceleration for gaofen-3 mission. *IEEE J Sel Top Appl Earth Obs Remote Sens* 2018;**11**(6):2115–26.
21. Wang G, Tan S, Guan C, Wang N, Liu Z. Multiple model particle filter track-before-detect for range ambiguous radar. *Chin J Aeronaut* 2013;**26**(6):1477–87.
22. Rohling H. *New CFAR-processor based on an ordered statistic*. International Radar Conference; 1985.
23. Durand R, Thirionlefevre L, Ginolhac G, Forster P. SAR processor based on a CFAR signal or interference subspace detector matched to a man-made target detection in a forest. IEEE international conference on Acoustics; 2007 Apr 15–20; Honolulu, USA. Piscataway: IEEE Press; 2007. p. 293–6.
24. Tian MH, Wan SH, Yue LH. Ship detection in remote sensing images with complex sea surface background. *J Chin Comput Syst* 2008;**29**:2162.
25. Paek KH, Yao M. A review on the application of visual attention in target detection of remote sensing image. 2nd international conference on remote sensing, environment and transportation engineering; 2012 Jun 1–3; Nanjing, China. Piscataway: IEEE Press; 2012.
26. Chen X, Huoa H, Taoa F, Lib D, Lia Z. A computational method to emulate bottom-up attention to remote sensing images. *XXI congress international society for photogrammetry and remote sensing* (ISPRS); 2008 Jul 3–11; Beijing, China: ISPRS; 2008. p. 244.
27. Liu C, Zhang D, Zhao X. Multitask saliency detection model for synthetic aperture radar (SAR) image and its application in SAR and optical image fusion. *J Electron Imaging* 2018;**27**(2):023026.
28. Ni J, Zhang Q, Yang Q, Luo Y, Sun L. Directional feature: A novel feature for group target detection in high resolution SAR images. *Remote Sens Lett* 2017;**8**(8):713–22.
29. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 2007;**9**(1):62–6.
30. Ai JQ, Yang XZ, Song JT, Dong ZY, Jia L, Zhou F. An adaptively truncated clutter-statistics-based two-parameter CFAR detector in SAR imagery. *IEEE J Oceanic Eng* 2018;**43**(1):267–79.
31. Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned salient region detection. *IEEE conference on computer vision & pattern recognition*; 2009 Jun 20–25; Miami, USA. Piscataway: IEEE Press; 2009. p. 1597–640.
32. Hou X, Zhang L. Saliency detection: A spectral residual approach. *2007 IEEE conference on computer vision and pattern recognition*; 2007 Jun 17–22;Minneapolis, USA. Piscataway: IEEE 1112 Press; 2007; p. 1–8.
33. Vanrell M, Otazu X, Parraga CA. Saliency estimation using a non-parametric low-level vision model. *Computer vision & pattern recognition*; 2011 Jun 20–25; Colorado Springs, USA. Piscataway: IEEE Press; 2011. p. 433–40.
34. America P, Milner R, Nierstrasz O, Tokoro M, Yonezawa. What 1119 is an object?*Workshop on object-based concurrent computing*, 1991 1120 Jul 15–16. London, UK: Springer-Verlag; 1991. p. 257–64.
35. Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2017 Jul 21–26; Honalulu, USA. Piscataway: IEEE Press; 2017. p. 7263.
36. Wang GH, Tan SC, Guan CB, Wang N, Liu ZL. Multiple model particle filter track-before-detect for range ambiguous radar. *Chin J Aeronaut* 2013;**26**(6):1477–87.