electronics

MDPI

Article

Military Object Real-Time Detection Technology Combined with Visual Salience and Psychology

Xia Hua [1,*], Xinqing Wang [1], Dong Wang [1,2], Jie Huang [1] and Xiaodong Hu [1]

[1] College of Field Engineering, PLA Army Engineering University, Nanjing 210007, China; wwwxxxqqq@126.com (X.W.); dyhkxydfbb@163.com (D.W.); huangjie051501@126.com (J.H.); hxd3281008@163.com (X.H.)

[2] Second Institute of Engineering Research and Design, Southern Theatre Command, Kunming 650222, China

* Correspondence: huaxia120888@163.com; Tel.: +86-025-8082-1050

Received: 6 August 2018; Accepted: 19 September 2018; Published: 25 September 2018

Abstract: This paper presents a method of military object detection through the combination of human visual salience and visual psychology, so as to achieve rapid and accurate detection of military objects on the vast and complex battlefield. Inspired by the process of human visual information processing, this paper establishes a salient region detection model based on double channel and feature fusion. In this model the pre-attention channel is to process information on the position and contrast of images, and the sub-attention channel is to integrate information on primary visual features first and then merges results of the two channels to determine the salient region. The main theory of Gestalt visual psychology is then used as the constraint condition to integrate the candidate salient regions and to obtain the object figure with overall perception. After that, the efficient sub-window search method is used to detect and filter the object in order to determine the location and range of objects. The experimental results show that, when compared with the existing algorithms, the algorithm proposed in this paper has prominent advantages in precision, effectiveness, and simplicity, which not only significantly reduces the effectiveness of battlefield camouflage and deception but also achieves the rapid and accurate detection of military objects, thus promoting its application prospect.

Keywords: salience detection; Gestalt visual psychology; human vision; machine vision; military object; adaptive

1. Introduction

In recent decades, wars depend more and more on advanced technology, and as a result, the patterns of warfare have changed from mechanized warfare to information warfare, which has become the main form of modern warfare. Rapid, efficient, and accurate detection of military objects for the purpose of accurate attacks is not only an indispensable demand for modern warfare, but also a crucial element for the improvement of early strategic warning systems and missiles [1].

Object detection is the basis of object tracking and recognition. The quality of detection results plays a decisive role in subsequent operations. At present, the commonly-used object detection methods mainly include several traditional ones, such as feature matching method, background modeling method, threshold segmentation method, methods based on depth learning, as well as methods based on visual salience. The feature matching method in the traditional detection algorithms (such as [2–4]) has high detection precision and accuracy, but it has low autonomy and low calculation efficiency, and its objects need to be initialized manually. Background modeling methods (such as [5–7]) can achieve automatic segmentation of objects and backgrounds, but the establishment and update of models is time-consuming and dynamic backgrounds will interfere with the results. Threshold segmentation methods (such as [8,9]) are convenient and efficient for situations with simple

backgrounds and prominent objects, but the detection effect under complex circumstances is not satisfactory. To sum up, traditional detection algorithms have limitations, which make it difficult to meet the needs of complex and diverse scenarios in real life. Moreover, these methods are subject to manual interference and their adaptive abilities are far from being enough.

The in-depth learning-based detection algorithms (such as [10–12]) can be applied to a variety of detection scenarios, since they are flexible and convenient for modeling on the one hand, and are highly diversified for the detection and identification of various types of objects on the other hand. This is the reason why they have been applied to lots of tasks, such as the monitoring and identification of vehicles and pedestrians. However, the detection effects of such algorithms depend too much on the construction of data sets, particularly the large data sets and manually labeled data sets, which means that it needs lots of computational resources.

The human eye's visual attention mechanism enables the visual system to extract the most interesting region from the huge image data, thus greatly improving the efficiency of data processing. Along with the development of neuropsychology and neuroanatomy, the visual attention mechanism has gradually become a hot topic in computer vision researches, and thereby attracted the attention of many scholars. So far, the existing visual attention models can be divided into visual attention prediction models and saliency region detection models according to their functions. The former ones are largely utilized to predict the amount of attention that is paid by human eyes to each pixel in the image, while the latter ones, as the major focus of this paper, are largely utilized to detect those salient objects in the image. Visual saliency is an object detection algorithm model that simulates human eyes, which can quickly focus on a particular region. It is mainly divided into a top-down saliency detection model and a bottom-up saliency detection model. The former one regards salient object detection as a learning problem, and actively searches for the required object salient map just as the tasks require. The latter is a salient map acquisition model designed to imitate the instinctive response of human beings to the scene. For a long time, many researchers have put forward various methods to obtain significant object, such as GBVS [13], FT [14], RC [15], CHM [16], DSR [17], etc. Apart from the above-mentioned methods and their improved versions, many new saliency detection methods while using depth learning have emerged during the past two years, such as ds SOD [18], RFCN [19], and DRFI [3], whose principles are to generate saliency maps by the construction and training of neural networks.

Gestalt psychology advocates the study of direct experience and behavior, emphasizes the integrity of experience and behavior, believes that the whole is not equal to and greater than the sum of parts, and advocates the study of psychological phenomena with the view of the dynamic structure of the whole. Reference [20] introduces an experimental paradigm to selectively probe the multiple levels of visual processing that influence the formation of object contours, perceptual boundaries, and illusory contours. Reference [21] presents psychophysical data derived from three-dimensional (3D) percepts of figure and ground that were generated by presenting two-dimensional (2D) images that are composed of spatially disjoint shapes that pointed inward or outward relative to the continuous boundaries that they induced along their collinear edges. The experiments of reference [22] reported herein probe the visual cortical mechanisms that control near-far percepts in response to two-dimensional stimuli. Figural contrast is found to be a principal factor for the emergence of percepts of near versus far in pictorial stimuli, especially when the stimulus duration is brief.

Due to military regulations on classified information, few systematic researches in this field have been carried out at home and abroad, and as for those researches that have been done recently, they lack a framework dedicated to military object detection tasks. When compared with conventional object detection tasks, weapons and personnel on battlefields will be disguised to a certain extent in order to improve the survivability probability of weapons and equipment and to enhance the survivability of personnel. During non-war time, military objects will also be disguised for the sake of classified information regulations on military facilities and equipment. Thus, camouflage, together with complex and changeable battlefields, actually makes it more difficult to detect military objects.

Taking into account the characteristics and requirements of military object detection tasks, this paper, highlighting the imitation of human visual perception mechanism, proposes a military object detection framework that combines human visual salience with visual psychology, and focuses its analyses and researches on the following five aspects:

(1) establishing a data set dedicated to military object detection to ensure that the data is sufficient and representative so as to compare and verify the validity of the model;
(2) imitating the human eye vision adaptive adjustment system, thus proposing a new method for image adaptive enhancement is proposed in order to highlight the object, weaken the background, and suppress interference;
(3) establishing a saliency region detection model based on double channel and feature fusion, after being inspired by how human visual information is processed; and,
(4) applying Gestalt's main theory on visual psychology as a constraint condition to integrate the obtained candidate salient regions and thus to generate a salient map with overall perception;
(5) proposing a new efficient sub-window search method to detect and screen objects and to determine the region where the objects are located on the other hand.

2. Our Model

The part of the cerebral cortex that is mainly responsible for processing visual information is the visual cortex, which includes a primary visual cortex (V1, also called "star irate cortex") and an extra cortex (such as V2, V3, etc.). As the first area to perform visual processing, V1 mainly receives electrical signals that are related to appearance perception, and the response results are further transmitted to higher-level visual cortex areas, such as V2 for processing [23]. The (a) [24] in Figure 1 shows the hierarchical structure of the cerebral visual cortex. Inspired by the visual cortex structure and Gestalt visual psychology, this paper established a three-layered spatial object detection model: the local salient regions of the object can be quickly detected, and then the detection results are simplified and fused layer by layer, thus making it a whole object that is perceptually easy to detect and process. The (b) in Figure 1 shows the hierarchical structure of our model.

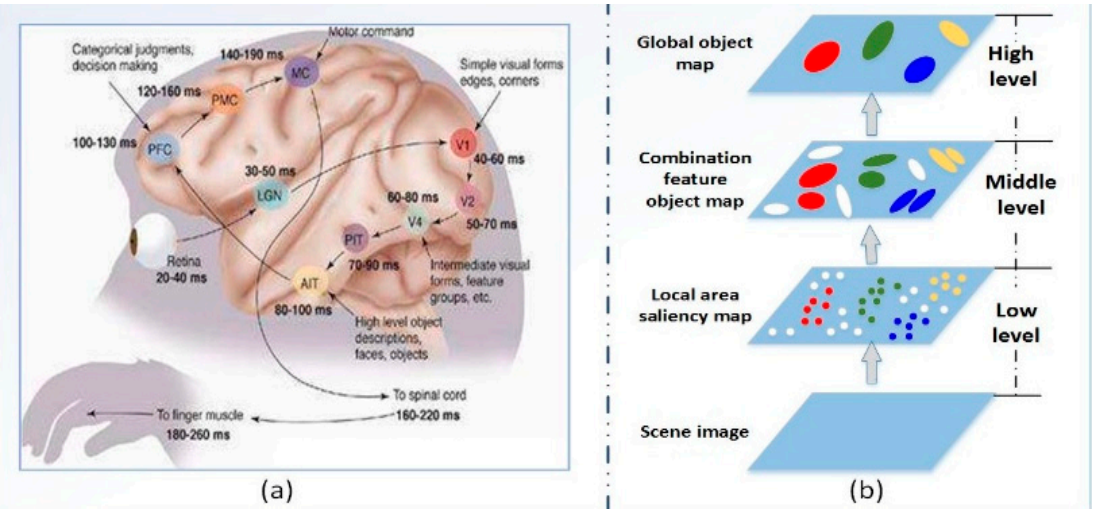


Figure 1. The hierarchical structure of the cerebral visual cortex and the hierarchical structure of our model. (**a**) The hierarchical structure of the cerebral visual cortex, [24]; (**b**) The hierarchical structure of our model.

The overall structure of our model proposed in this paper is shown in the Figure 2. The given input image that I can be dealt with, as follows:

(1) the image is effectively enhanced by a method based on human eye vision fusion [25] (yellow border section);

(2) a saliency region detection model based on dual channel and feature fusion (red border section) is established, wherein the pre-attention channel processing obtains the pre-attention saliency map F_{pre}, the sub-processing channel processing obtains the sub-attention saliency map F_{sub}; then, the two channel detection results are fused to obtain the final saliency map F_{final}; and,

(3) using Gestalt's main theory of visual psychology as a constraint condition (blue border section), the candidate salient regions that are obtained are integrated to achieve an object segmentation result map F with overall perception. Then, the object (blue border section) is detected and screened by using an efficient sub-window search method to determine the region where the object is located.

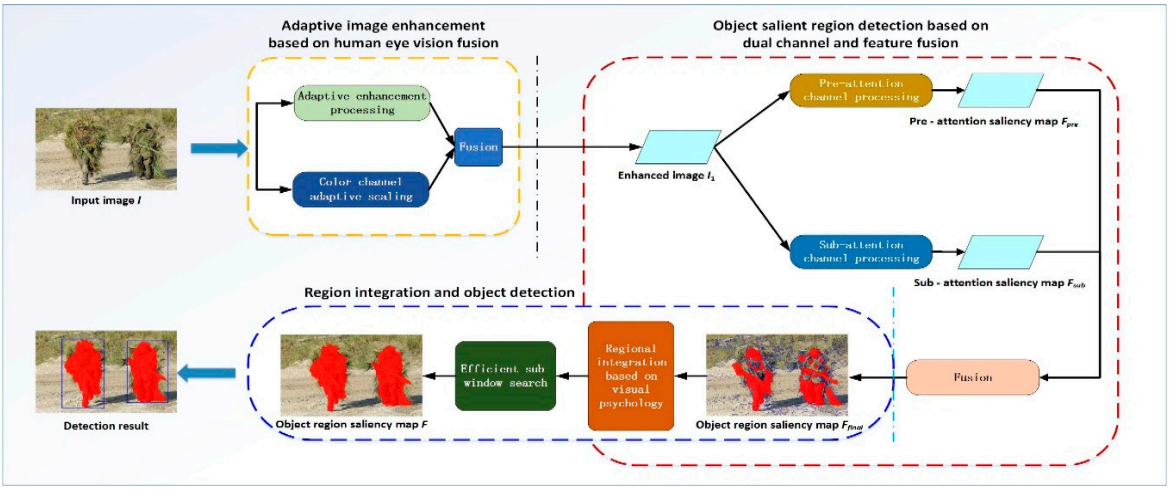


Figure 2. Illustration of our proposed network architecture.

3. Salient Region Detection Based on Double Channel and Feature Fusion

3.1. Human Visual Perception and Processing System

After long-term evolution and development, human eyes have formed an exquisite visual perception and processing system, which can quickly filter, process, and analyze the surrounding visual information. Neurobiology research shows that, in the process of perception of visual information, the visual information that is received is mainly analyzed and processed in two kinds of neural pathways: ventral channel and dorsal channel. Ventral channel, also called "What Channel", is mainly responsible for the cognition of object information and it can identify features, such as shape and color. The dorsal channel, also called, "Where Channel", is responsible for the movement and spatial information like the movement and position of objects [26], as can be shown in Figure 3.

After long-term evolution and development, human eyes have formed an exquisite visual perception and processing system, which can quickly filter, process, and analyze the surrounding visual information. Neurobiology research shows that, in the process of perception of visual information, the visual information that is received is mainly analyzed and processed in two kinds of neural pathways: ventral channel and dorsal channel. Ventral channel, also called "What Channel", is mainly responsible for the cognition of object information and it can identify features, such as shape and color. The dorsal channel, also called, "Where Channel", is responsible for the movement and spatial information like the movement and position of objects [26], as can be shown in Figure 3.

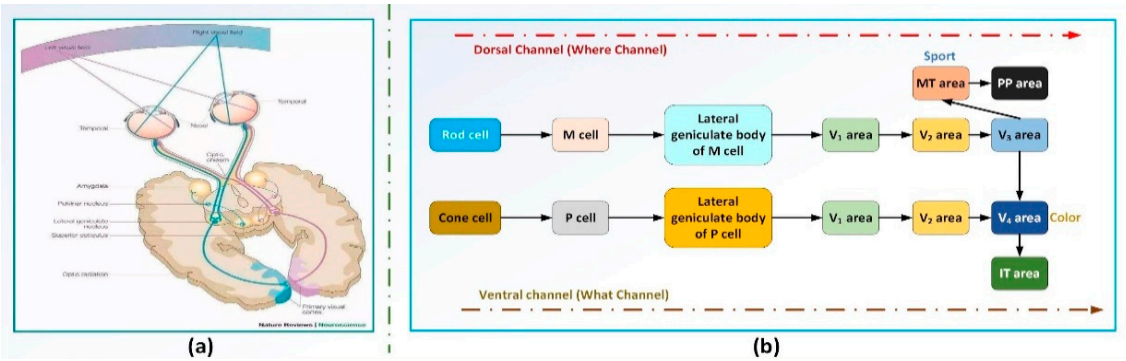


Figure 3. Human visual perception and processing system. (**a**) A diagram of information transmission of the visual channel [27]; (**b**) A schematic diagram of information transmission of the visual channel.

3.2. Salient Region Detection Model Based on Dual Channel and Feature Fusion

In the process of forming visual attention, there are mainly two kinds of attention selection mechanisms. One is bottom-up attention and is mainly driven by the underlying data. Under unconscious control, this mechanism belongs to a lower level of visual cognition and it has a faster processing speed. The other one, also called task-driven visual attention, is top-down attention and is greatly influenced by prior knowledge, tasks, and expectations. It emphasizes the dominant role of consciousness and it belongs to the high-level information processing and perception. In the process of visual information processing and analyzing, the bottom-up and top-down visual attention processes coordinate with each other to realize effective perception of external environmental information [28].

Inspired by human visual perception and processing, this paper proposes a new saliency region detection model that is based on dual channel and feature fusion, as shown in Figure 4. This model establishes a pre-attention channel to simulate the "Where Channel" of the visual system to process information on the position and contrast information of images. Then, it uses a sub-attention channel to simulate the channel, integrate the primary visual feature information in the image, and carry out significant calculation. Finally, the pre-attention channel and the post-attention channel are significantly integrated, and the gaze region is determined according to the final saliency measurement result.

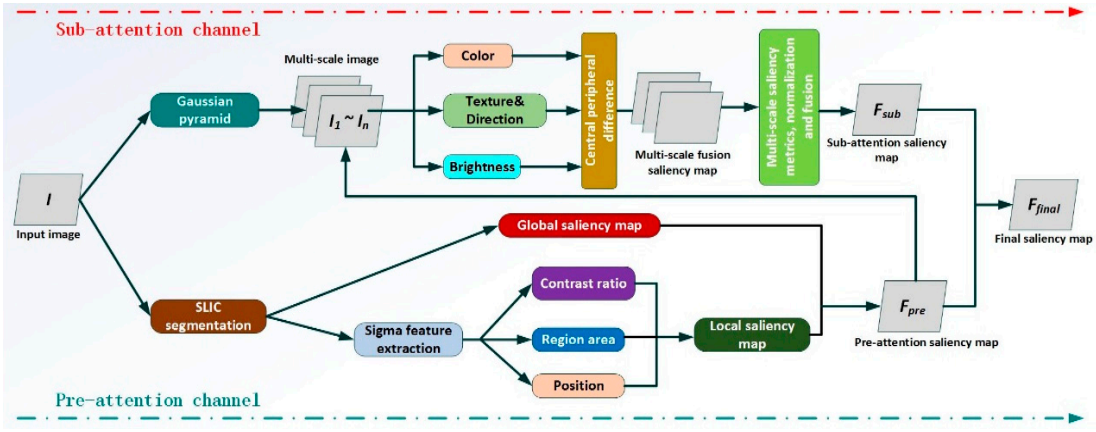


Figure 4. The saliency region detection model based on dual channel and feature fusion. Our model mainly includes two parts, the sub-attention channel (indicated by the red dotted line) and the pre-attention channel (indicated by the green dotted line).

This model mainly includes the following three aspects:

(1) Pre-attention channel processing. First, the SLIC super pixel segmentation [29] is performed on the color image, and then sigma features of each region are extracted [30]. While considering the influence of neighborhood contrast, background contrast, spatial distance and region size,

a local saliency map is generated, and then the over-all density estimation of each region is extracted to construct a global saliency map. Finally, the local and global saliency maps are fused by exponential weighting to construct a saliency map F_{pre}.

(2) Sub-attention channel processing. Firstly, multi-resolution processing is carried out on the input image by using a variable-scale Gaussian function to establish a Gaussian pyramid [31]. Then, the color space of the image is transformed according to the color antagonism mechanism. After that, information on such features as color, brightness, and texture orientation of potential objects' image region is fed back by the pre-paid attention path to be extracted on multiple scales, which leads to the generation of the corresponding feature map by using the central peripheral difference [32] operation. Finally, the saliency of each primary visual feature is measured through inter-layer Gaussian difference processing [33], and the saliency of each feature is integrated between layers, and the saliency map F_{sub} is calculated by combining the weighted summation method.

(3) Normalizing the pre-attention saliency map and the post-attention saliency map, merging the pre-attention channel and the sub-attention channel, and identifying a salient object region according to the total saliency map F_{final}.

Because the processing time of the pre-attention channel is much lower than that of the sub-attention channel, the results of the pre-attention channel can effectively guide the detection of the sub-attention channel, rendering the two channels to operate side by side.

3.2.1. Pre-Attention Channel Based on Fused Saliency Map

SLIC (simple linear iterative clustering), which is a simple linear iterative clustering, is a simple and easy-to-implement algorithm proposed in 2010. It converts color images into CIELAB color space and five-dimensional feature vectors in the X and Y coordinates. Then, it constructs distance metrics for five-dimensional feature vectors and performs local clustering [30]. The SLIC algorithm can generate compact, nearly uniform super pixels, and it has a high comprehensive evaluation in terms of computation speed, object contour retention, and super pixel shape, which is in line with the desired segmentation effect.

Since the regional covariance can naturally fuse multiple related features, the covariance calculation itself has filtering ability and high efficiency. Therefore, this paper uses the regional covariance [34] feature to perform image local saliency detection. The following five features are extracted for each image pixel: the image grayscale, the first and second degree norms of the x and y directions, so each pixel is mapped to a fi e-dimensional feature vector:

$$F(x,y) = \left[I(x,y), \left| \frac{\partial I(x,y)}{\partial x} \right|, \left| \frac{\partial I(x,y)}{\partial y} \right|, \left| \frac{\partial^2 I(x,y)}{\partial x^2} \right|, \left| \frac{\partial^2 I(x,y)}{\partial y^2} \right| \right]^{\mathrm{T}} \tag{1}$$

Here, I is the gray level of the image and the image gradient is calculated according to the referencing material [35]. The covariance matrix of the region R is calculated as Equation (2).

$$\begin{cases} C_{ov\,R} = \frac{1}{N(R)} \sum_{i=1}^{n} (F_i - \mu)^{\mathrm{T}} (F_i - \mu) \\ \mu = \frac{1}{N(R)} \sum_{i=1}^{n} F_i \end{cases} \tag{2}$$

In the equation, μ is the mean of the region feature vectors and N (R) represents the number of pixels in the region R. In order to enable the regional covariance to better reflect the image region characteristics and to facilitate the calculation of similarity, C is the covariance matrix of $d \times d$ dimension, and the Sigma feature is introduced [36]:

$$s_i = \begin{cases} \alpha\sqrt{d}L_i, if 1 \le i \le d \\ -\alpha\sqrt{d}L_i, if 1+d \le i \le 2d \end{cases} \tag{3}$$

In the equation, L_i is the i-th column of the matrix L, $A_{rea} = LL^{\mathrm{T}}$, α is a coefficient, the mean of the d-dimensional vector is introduced, and the Sigma eigenvector of $\boldsymbol{A_{rea}}$ enhancement is Equation (4):

$$\psi_{A_{rea}} = (s_1 + \mu, s_2 + \mu, \cdots, s_{2d} + \mu)^{\mathrm{T}} \tag{4}$$

The local saliency of the region R_i is defined as the spatial distance weighted average of the enhanced Sigma features of the region R_i and its neighboring regions, as shown in Equation (5):

$$S_c(R_i) = \frac{1}{mK} \sum_{R_j \in Nb} \exp\left(-\frac{\|R_i - R_j\|^2}{{\sigma_1}^2}\right) \cdot D\left(\psi_{R_i}, \psi_{R_j}\right) \tag{5}$$

In the equation, R_j belongs to the neighborhood of the region R_i; m is the number of neighborhood regions; K is the normalization factor; the sum of the spatial distance weighting coefficients is guaranteed to be 1; $\|R_i - R_j\|$ is the Euclidean distance between the centers of the two regions. What is more, σ_1 controls the effect of inter-region distance on local saliency, and the larger the σ_1, the greater the influence of the distant block on the saliency of the current block. ψ_{Ri} represents the enhanced Sigma characteristic of the region R_i, and $D\left(\psi_{R_i}, \psi_{R_j}\right)$ is the Euclidean distance of ψ_{Ri} and ψ_{Rj}. In general, when human eyes observe the surrounding information, more attention will be paid to the central area in the field of view. When the two adjacent areas are compared, the large area should have a greater influence on the current area. At the same time, a significant object area has not only a high local contrast, but also a background area. The domain differences are salient [30]. When the effects of neighborhood contrast, background contrast, spatial distance, and region size are taken into consideration, the local saliency of the improved region R_i is Equation (6):

$$S_l(R_i) = \frac{1}{mK} \sum_{R_j \in Nb\&bg} \exp\left(-\frac{\|R_i - R_j\|^2}{{\sigma_1}^2}\right) \cdot \frac{N(R_j)}{N(I)} D\left(\psi_{R_i}, \psi_{R_j}\right) \tag{6}$$

In the equation, $N(R_j)$ represents the number of pixels in the region R_j, $N(I)$ represents the number of pixels of the image, and R_j belongs to the neighborhood of the current region and the boundary region of the image. The probability that the gray value of each region appears might indicate the global saliency of the region, and the object region with a low probability of occurrence means more salient, and conversely, it may be the background region. Therefore, it is possible to use the kernel density estimation of the entire image region feature to calculate the overall saliency, specifically:

$$\begin{cases} S_g(R_i) = \frac{\sum_{i=1}^{M}\sum_{j=1}^{M}\kappa\left(I_{R_i} - I_{R_j}\right)}{\sum_{j=1}^{M}\kappa\left(I_{R_i} - I_{R_j}\right)} \\ \kappa\left(I_{R_i} - I_{R_j}\right) = \exp\left(\frac{-\|I_{R_i} - I_{R_j}\|^2}{{\sigma_2}^2}\right) \end{cases} \tag{7}$$

Here, $\kappa(x)$ is a Gaussian kernel density function, I_{Ri} represents the average gray level of the region, and m is the number of image regions. Given that the local saliency is usually better than the overall saliency and that the exponential function can increase the importance of local saliency, the equation of the pre-attention channel saliency map that was obtained by combining local and global saliency maps is designed, as Equation (8):

$$F_{pre} = S_g(R_i) \times \exp(\sigma_3 \times S_l(R_i)) \tag{8}$$

Among them, σ_3 controls the importance of a local saliency map. The fusion saliency map can ensure that the object area is both locally salient and globally salient, which is beneficial to reduce the influence of background clutter in subsequent object detection and segmentation. There are altogether seven parameters of the channel saliency map: *ns* (num super pixels), *CP* (comparison), σ_1, σ_2, σ_3, *C*, and *ra* (ratio). Among them, *ns* and *CP* are parameters of SLIC method, *ns* is the number of super pixels, and the smaller the value is, the larger the super pixel block there will be. *CP* indicates the shape of the super-pixel, and the smaller the value is, the higher consistency there is between the super pixel block and the region block boundary. *ns* and *CP* need to be adjusted according to different images. $\sigma_1 = 3$, $\sigma_2 = 10$, $\sigma_3 = 6$, $C = 2000$, $ra > 0.5$.

3.2.2. Sub-Attention Channel Based on Gaussian Pyramid and Feature Fusion

The retina of the human eye samples the image information unevenly. The resolution of the central region is higher, while the resolution of the peripheral region is lower. The visual perception and processing system uses the interaction between the receptive field and the integrated field to represent image information in multi-resolution. In many cases, features that are not easy to see or acquire in one scale can be easily found or extracted in another scale [37]. According to this mechanism, before each primary visual feature is processed, multi-scale representation of the image is required, and then the central peripheral difference is calculated through the interlayer subtraction operation to generate the corresponding feature map. We use the Gaussian function with variable parameters to process the input image to obtain the Gaussian pyramid of the image. The images of each layer are obtained by Gaussian filtering with different scales instead of changing the resolution of the input image itself. The initial value of standard deviation σ of the Gaussian filter used in this paper is 3 and the image is processed with $\sqrt{2}$ as multiple to generate images of different scales. Defining the pixels of the input image as $I(i, j)$ and the number of layers of the Gaussian pyramid as k ($k = 0, 1, 8$), the image of the k-th layer can be derived from the Equation (9):

$$I_k(i,j) = w(m,n)_k \otimes I(i,j) \tag{9}$$

Here, $w(m, n)_k$ is the Gaussian function used in the k-layer image and $\otimes$ represents convolution operation. After multi-scale representation of the image, the color, brightness and direction characteristics of each layer of the image need to be extracted, and then the feature map of the image at each scale can be obtained. The human eye is able to detect significant objects with various characteristics, as shown in the Figure 5.

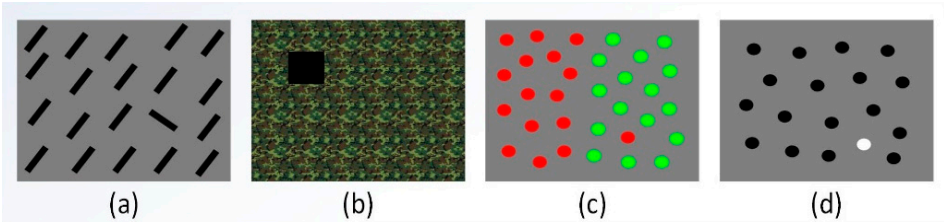


Figure 5. The Schematic diagram of saliency of various features. In the figure, (**a**) is salient for the direction feature; (**b**) is salient for the texture feature; (**c**) is salient for the color feature; and, (**d**) is salient for the luminance feature.

The brightness characteristics of the images of each layer of the Gaussian pyramid can be obtained by Equation (10):

$$I(k) = \frac{(r(k) + g(k) + b(k))}{3} \tag{10}$$

In the equation, $r(k)$, $g(k)$, and $b(k)$ represent red, green, and blue color channels, respectively. In some scenes, brightness features are used to enhance saliency, which means that areas with high

brightness have strong saliency. However, experiments show that, in battlefield environments, the brightness of objects is generally lower than that of the environment, and therefore brightness features are used to play an inhibitory role in the saliency measurement in this paper. For color features, it is feasible to simulate the color perception process of the human eye according to the color antagonism theory, while using the red-green (RG) and blue-yellow (BY) color models [33] for calculation. In the k-th layer image, the RG color feature $M_{RG}(k)$ and the BY color feature $M_{BY}(k)$ are calculated by Equation (11).

$$\begin{cases} M_{RG}(k) = \frac{r(k)-g(k)}{\max(r(k),g(k),b(k))} \\ M_{BY}(k) = \frac{b(k)-\min(g(k),r(k))}{\max(r(k),g(k),b(k))} \\ M_{color}(k) = \max(M_{RG}(k), M_{BY}(k)) \\ M_{RG}(k) = M_{BY}(k) = 0, if \max(r(k),g(k),b(k)) < 0.1 \end{cases} \tag{11}$$

Since the Gabor function can achieve better results in acquiring the directional features, a two-dimensional Gabor filter is used to extract the texture and directional features of the image. The mathematical expression of the two-dimensional Gabor function [33] is shown in the Equation (12).

$$g(x, y, \lambda, \theta, \psi, \sigma, \gamma) = \exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) \exp(i(2\pi \frac{x'}{\lambda} + \psi)) \tag{12}$$

In the equation, x, y represent pixel coordinates, $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$, λ represents the sine function wavelength; θ represents the direction of the kernel function; ψ represents the phase offset; σ represents the Gaussian function Standard deviation; and, γ represents the ratio of width to height of space. Analysis of the shape characteristics of the military object itself leads to the finding that the main texture of the military object is a vertical and horizontal straight line, as well as a circular or circular curve. To this end, we designed three Gabor convolution kernels (as shown in Figure 6), including vertical-direction kernel, a horizontal-direction kernel, and a symmetric circular kernel, to extract the texture and directional features of the image through convolution operation.

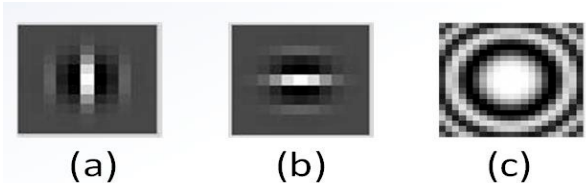


Figure 6. Three forms of Gabor convolution kernel. (**a**) is a vertical convolution kernel; (**b**) is a horizontal convolution kernel; and, (**c**) is a circular convolution kernel.

The central peripheral difference refers to the difference between images of different layers of the Gaussian pyramid, which is realized by subtraction of images without neglecting scale. For the peripheral layer with a smaller number of pixels, interpolation processing is needed so that the number of pixels is equal to the corresponding central layer, and subtraction operation is performed for each pixel point [33]. Then, the luminance characteristic map is derived from the Equation (13).

$$M_{I(c,c+s)} = |I(c) - I(c+s)| \tag{13}$$

Here, c and $c + s$ represent the number of layers of the image, $c \in \{3,4,5\}$, $s \in \{3,4\}$, which represents the subtraction operation between images of different layers. The color feature map Mcolor can be generated, as Equation (14):

$$\begin{cases} M_{RG(c,c+s)} = |(R(c) - G(c)) - (G(c+s) - R(c+s))| \\ M_{BY(c,c+s)} = |(B(c) - Y(c)) - (Y(c+s) - B(c+s))| \\ M_{color(c,c+s)} = \max(M_{RG(c,c+s)}, M_{BY(c,c+s)}) \end{cases} \tag{14}$$

In the equation, M_{RG} $(c, c + s)$ and M_{BY} $(c, c + s)$, respectively, represent the color feature maps of the red-green channel and the blue-green channel. The direction feature map is calculated according to the Equation (15).

$$M_{ori(c,c+s,\theta)} = |(c,\theta) - (c+s,\theta)| \tag{15}$$

θ is the positive direction of Gabor filter output, (c,θ) represents the directional feature map in θ direction when the scale space is c. After obtaining the feature maps, the saliency of each feature map can be measured by the following Equations (16) and (17).

$$DOG(x,y) = \frac{{c_{ex}}^2}{2\pi{\partial_{ex}}^2}\exp\left[-\frac{x^2+y^2}{2{\partial_{ex}}^2}\right] - \frac{{c_{inh}}^2}{2\pi{\partial_{inh}}^2}\exp\left[-\frac{x^2+y^2}{2{\partial_{inh}}^2}\right] \tag{16}$$

$$N\left(M_{(c,c+s)}\right) = \left(M_{(c,c+s)} + M_{(c,c+s)} * DOG - C\right) \tag{17}$$

Here, $DOG(x,y)$ represents a Gaussian difference function; $N(M_{(c,c+s)})$ is a saliency measure function; ∂_{ex} and ∂_{inh} represents the bandwidths of stimulation and suppression parameters, which are set to be 0.02 and 0.25; and, the three constants of c_{ex}, c_{inh}, and C were set to be 0.5, 1.5, and 0.02. Using the above-set parameters for 10 iterations, a feature saliency map can be produced. The feature saliency maps of the different features of each layer image also need to be integrated to obtain saliency maps of interlayer features in terms of brightness, color, and texture direction respectively. $\oplus$ represents the operation of expanding the matrix and summing it up one by one.

$$\begin{cases} F_I = N\left(\bigoplus_{C=2}^{4}\bigoplus_{S=2}^{4} N\left(M_{i(c,c+s)}\right)\right) \\ F_c = N\left(\bigoplus_{C=2}^{4}\bigoplus_{S=2}^{4} N\left(M_{color(c,c+s)}\right)\right) \\ F_o = N\left(\sum_{\theta}\bigoplus_{C=2}^{4}\bigoplus_{S=2}^{4} N\left(M_{o(c,c+s,\theta)}\right)\right) \end{cases} \tag{18}$$

The traditional algorithm linearly superimposes the multi-feature saliency map, while Gestalt does not take the whole as the sum of each part. Therefore, we use the non-linear combination of multi-feature images and using the minimum F–norm [37] to constrain and obtain the most competitive local saliency region. The F-norm of the matrix A is expressed by Equation (19), and the nonlinear combination mode of the characteristic image expressed by Equation (20) is determined by the parameter $\theta_1 = (a, b, c_1, e, g, h)$. Where a, b, c_1 are simplified parameters, a, b, $c_1 \in \{1, 2, 3\}$, e, g, h are combined parameters, $e, g, h \in \{-1, 1\}$. The combination parameter value of -1 indicates that the feature region under this attribute has a negative effect on the extraction of the best salient region, while the value of 1 indicates that the feature region under this attribute has a positive effect on the extraction of the best salient region. As shown in the Equation (21), the nonlinear combination parameter of the salient region corresponding to the feature is obtained by using the minimum F-norm. This generates the gross salient map of the post-processing path, which has the strongest salient and ensures that the salient region corresponding to the nonlinear combination feature is sufficiently sparse.

$$\|A\|_F = \sqrt{\sum_i\sum_j {a_{ij}}^2} \tag{19}$$

$$F_{sub} = e * {F_I}^a + g * {F_c}^b + h * {F_o}^{c_1} \tag{20}$$

$$\theta_1 = \underset{\theta_1}{\operatorname{argmin}}\|F_{sub}(\theta_1)\|_F \tag{21}$$

3.2.3. Gross Saliency Map

The salient map of the pre-attention channel and the sub-attention channel is integrated by Equation (22).

$$F_{final} = \eta_1 \frac{F_{pre}}{F_{pre}^{\max}} \circ \eta_2 \frac{F_{sub}}{F_{sub}^{\max}} \tag{22}$$

In the equation, $F_{pre}^{\max}$ and $F_{sub}^{\max}$ are the maximum values of F_{pre} and F_{sub}, η_1 and η_2 are weight parameters. "$\circ$" represents an arithmetic symbol. Based on experimental tests, we use "+" as the arithmetic symbol to perform non-maximum suppression operations. Through experimental statistics, the conclusion can be drawn: the best fusion effect can be obtained when weights $\eta_1 = 0.326$, $\eta_2 = 0.674$.

4. Regional Integration of Visual Overall Perception Based on Gestalt

4.1. Efficient Sub-Window Search

Based on the above-explained methods, a saliency map of the image comes into being in which the value of pixels is the assessment of the saliency of the input pixel. The higher the pixel value is, the stronger the saliency can be. Accordingly, this paper proposed an efficient sub-window search algorithm, and the result of high efficient sub-window search is the encircling box that holds the object with the specific algorithm procedure demonstrated in Figure 7.

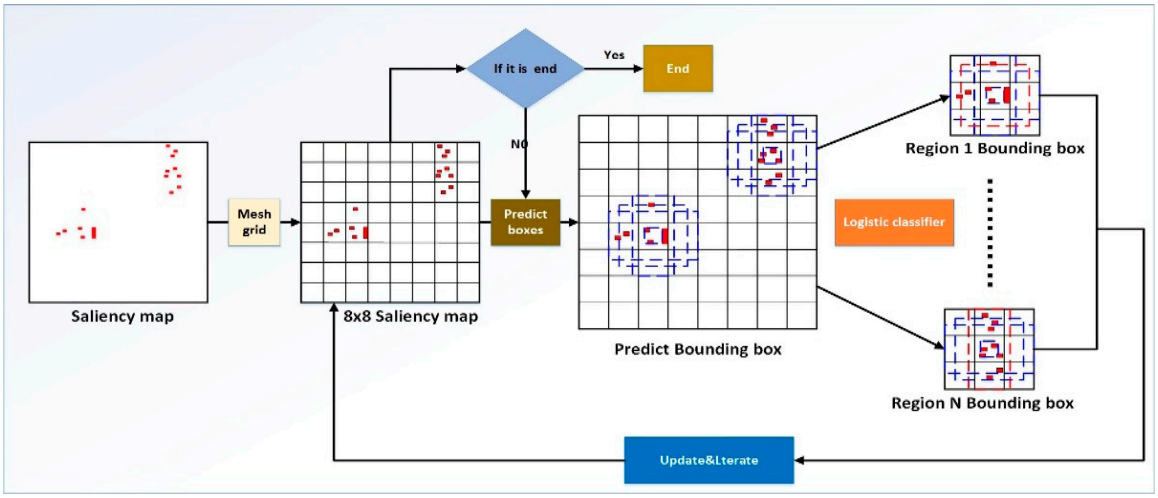


Figure 7. The efficient sub window search algorithm.

The saliency map is divided into an equal rectangle area according to the 8 × 8 grid, and the sum of pixel values in each rectangle is counted and the threshold is set. The center block of each region is constituted by 3 × 3 rectangles that form the object screening area. In the same object screening region, if multiple rectangular regions meet the requirements set by the pixel value, then the largest one is taken as the center of the region. If the center of the region is selected to be the large square edge, then the object screening area of the 3 × 3 can be formed by adding the small square of the same size blank. In the object screening area, the center of object screening area serves as the center, and four prediction bounding boxes with different width-height-length ratios are constructed according to different aspect ratios. Afterwards, the best object area bounding box is selected through the comparison of the construction index (pixel value, proportion and area) in each prediction bounding box. Thus, comes the total pixel value $s_{ump}x_i$ of rtg_i in the rectangular area of the saliency map after grid division.

$$s_{ump}x_i = \sum_{j \in rtg_i} px_j \tag{23}$$

The px_j represents the pixel value of the j pixel in the rtg_i area. The larger the $s_{ump}x_i$ value is, the greater the saliency of rtg_i in the rectangular area can be. Recognizing the saliency object area as the center accords with how human eyes' cognize the block domain. We choose the threshold of pixel value adaptively through the two order difference method to complete the initial screening of the regional center blocks. The two order difference can represent the changing trend of the discrete array and can be used to determine the threshold in a set of pixel values. A saliency map is detected by default to get 64 candidate regions, and at last, each candidate region gets one overall pixel value $s_{ump}x_i$ to represent the of the significant strength and weakness. Arrays of 64×1 can be obtained. The elements that are less than 0.1 are rounding off without any object and get an array of $n \times 1$, the C. Then we set the function $f(C_k)$ for estimating $s_{ump}x_i$ decreasing trend, as Equation (24).

$$f(C_k) = \frac{(C_{k+1} - C_k) - (C_k - C_{k-1})}{C_k}, k = 2, 3, \cdots, n-1 \tag{24}$$

The C_k of the maximum value of $f(C_k)$ is taken as the $s_{ump}x_i$ threshold of this saliency map. In order to reduce the computation of the region, to improve the efficiency and real-time performance of the algorithm, and to ensure that the selected area has a better tolerance, we form an object screening area with 3×3 rectangles centered in each area center block. If more than a rectangular area meets the requirements set by pixel threshold conditions in the same object screening area, then the one that takes the largest pixel value is chosen as the center of the region. We take the center point of the object screening area as the center, and predict six bounding boxes with fixed size according to the different width-length-height ratio. The area of the object screening area is S_Z, and the area of each prediction bounding box is shown by Equation (25).

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m-1}(k-1), k = 1, 2, \cdots, 5 \tag{25}$$

In the equation, $s_{min} = 0.1 \times S_Z$, $s_{max} = 0.7 \times S_Z$, $m = 5$, and different bounding lengths are given for different prediction bounding boxes.

$$a_r = \frac{W}{H}, a_r \in \left\{1, 2, 3, \frac{1}{2}, \frac{1}{3}\right\} \tag{26}$$

W and H represent the width and length of the box, respectively. The width and length corresponding to the bounding box are predicted to be $H_k = \sqrt{\frac{s_k}{a_r}}$, $W_k = \sqrt{a_r \cdot s_k}$. When $a_r = 1$, there is also a prediction bounding box [38] with a size $s'_k = \sqrt{s_k \cdot s_{k+1}}$ that has a total of 6 predicted bounding boxes. For any box b_l, we calculate the total pixel value $s_{ump}x_i$ in the box, the region area is S, and salient pixel proportion in the region P_p is:

$$p_p(b_l) = \frac{pn_1}{pn_2} \tag{27}$$

Pn_1 represents the total number of salient pixels (i.e., pixels with a pixel value greater than 0.1) in the b_l region, and Pn_1 represents the total number of pixels in the b_l region. That is, for every prediction bounding box, b_l has feature vector (x, y, $sumpx$, W, H, Pp), x and y, respectively, represent the Upper left corner coordinates of b_l. A simple logistic classifier is trained to evaluate the effectiveness of each prediction bounding box by using the artificially calibrated training samples. We divide the evaluation results into two categories: object bounding box and non-object bounding box. For input bounding boxes, b_l = (x, y, $sumpx$, W, H, Pp), logistic classifier [39] introduces weight parameter θ = (θ_1, θ_2, $\ldots$, θ_6), weighting the attributes in b_l, obtaining $\theta^T b_l$, and introducing the logistic function to get the function $h_\theta(b_l)$.

$$h_\theta(b_l) = \frac{1}{1 + e^{-\theta^T b_l}} \tag{28}$$

The probability estimation function can be further obtained as Equation (29):

$$P(Y|\ b_l;\ \theta) = \begin{cases} h_\theta(b_l);\ Y = 1 \\ 1 - h_\theta(b_l);\ Y = 0 \end{cases} \tag{29}$$

It means that the probability of a tag being Y is given when the test sample b_l and parameter θ are given. From the test sample set and the training sample set, we can obtain their joint probability density, that is, likelihood function:

$$\prod_{i=1}^{n} P\left(y^{(i)} \middle|\ b_l{}^{(i)};\ \theta\right) = \prod_{i=1}^{n} (h_\theta(b_l))^{y(i)} (1 - h_\theta(b_l))^{1-y(i)} \tag{30}$$

Maximize the likelihood function and find the appropriate parameter θ, and thereby comes the Equation (30):

$$\ell(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(b_l) + \left(1 - y^{(i)}\right) \log(1 - h_\theta(b_l)) \tag{31}$$

According to the Equation (31), the parameter θ is obtained by the gradient descent method. The parameter θ is first derived:

$$\frac{\partial}{\partial \theta_j} \ell(\theta) = (y - h_\theta(b_l)) b_{lj} \tag{32}$$

Update rule:

$$\theta_j = \theta_j + \alpha\left(\left(y^{(i)} - h_\theta\left(b_l{}^{(i)}\right)\right) b_{lj}{}^{(i)}\right) \tag{33}$$

Having evaluated the box selection effect of each bounding box by the logistic classifier, we can obtain a corresponding frame selection evaluation score for each prediction bounding box. After that, a non-maximum suppression is carried out to get the final prediction as the final object bounding box. After completing the selection of the object bounding box, we set the pixel values in the enlarged filter area to be 0, while avoiding the inefficiency that is caused by repeated selection. Then, we update the corresponding region of the Saliency map and judge whether all the significant object regions in the saliency map have been detected (pixels value in saliency map is 0 or not). If it is true then the test is completed, if not, then we need to go through the detection process again.

4.2. Regional Integration Based on Visual Overall Perception

Gestalt theory clearly points out that under the influence of eyes and brain, images are constantly organized, simplified and unified. Gestalt's organizational process is to selectively unify some elements together, and we can perceive that it is a complete unit [25]. This paper mainly applies the following Gestalt's main theory [40] as a constraint to make salient regional integration:

(1) simplification: excluding the unimportant part of the background from the image, only preserving the necessary components, such as the object, so as to achieve the simplification of the vision;
(2) the relationship between the subject and the background: the feature of the scene will affect the parsing of the subject and the background in the scene, when a small object (or color block) overlaps with the larger object, we incline to the view that small objects are the main body and the big object is the background;
(3) the global and the local: the whole of the experience organized by the perceptual activity, in nature, is not equal to the simple linear superposition of the part;
(4) closeness: referring to the situation in which individual visual unit is infinitely close so that they form a large and unified whole;
(5) closed-up: closed figures are often regarded as a whole; and,
(6) incomplete or irregular graphics: they are often regarded as a similar regular pattern.

The constraint condition 1 and 3 have been embodied in pre attention channel processing and post attention channel processing. The significant image in the prepaid channel is simplified by simplifying the parameters; in the post attention channel processing, the saliency region is constructed by nonlinear combination of the characteristic significant images. At the high level, according to the Gestalt constraint condition 4 and the constraint condition 5, we judge whether there is any intersection between the significant regional encircling boxes T_i obtained after the high validity sub-window search in total salience diagram F_{final}. Then, we merge them according to the nearest principle. The combination condition is that there is an overlapping part [40] in the two object encircling box regions, and the proportion of the overlapping region T_k in any region is above a certain threshold φ (satisfying fourth, fifth), such as Equation (34).

$$\begin{cases} T_i \cap T_j \neq \varnothing \\ \frac{T_k}{T_i} \geq \varphi \ \ or \ \ \frac{T_k}{T_i} \geq \varphi \end{cases} \tag{34}$$

For two significant regional bounding boxes $T_i(x_i,y_i,a_i,b_i)$ and $T_j(x_j,y_j,a_j,b_j)$, x,y,a,b are the upper left corner coordinates and the width and height of the regional encircling boxes. After merging, a new bounding box is generated for $T_k(x_k,y_k,a_k,b_k)$. Combination rules is indicated by Equations (35) and (36):

$$\begin{cases} x_k = \min(x_i, x_j) \\ y_k = \min(y_i, y_j) \end{cases} \tag{35}$$

$$\begin{cases} a_k = |x_i - x_j| + a, a = \begin{cases} a_i, x_j < x_i \\ a_j, x_i < x_j \end{cases} \\ b_k = |y_i - y_j| + b, b = \begin{cases} b_i, b_j < b_i \\ b_j, b_i < b_j \end{cases} \end{cases} \tag{36}$$

In order to prevent the mutual interference between the fused bounding boxes that reduces the accuracy and precision of the multi object detection results, the bounding boxes are processed in the left-to-right and top-to-bottom order. For the significant object area in the encircling boxes, the closed operation in the morphological operation is used to form the overall significant object area. The merging will stop on the condition that the area of the saliency region satisfies the needs that the salient region is larger than the background area or than the contact image boundary (satisfying the constraint condition 2).

5. Experience & Analysis

5.1. Experience Condition and Dataset

The experimental hardware conditions in this paper are DELL Precision R7910 (AWR7910) graphics workstation, with the processor Intel Xeon E5-2603 v2 (1.8 GHz/10 M) and software Matlab2015a. Because there is no military object dataset that can be directly used at home and abroad, this paper collects the original image through the network search engine for the military object detection task, and it processes the image by enriching, stretching, rotating, etc. to enrich the dataset. A military object detection dataset (MOD) is constructed according to the VOC dataset format standard. Each image in the dataset corresponds to a label that identifies the image name, the category of the military object, and the height and width of the circumscribed rectangle. The dataset consists of more than 20,000 images whose size is unified into 480 × 480 pixels.

In order to verify the advantages and superiority of the proposed algorithm, we selected the following five datasets to evaluate the detection results, namely MSRA-B [41], HKU-IS [42], KITTI [43], PASCALS [44], and ECSSD [45]. These datasets are available on the Internet, and each contains a large number of images and well-divided annotations, which is why they have been widely used in recent years. MSRA-B contains 5000 images from hundreds of different categories, and due to its diversity

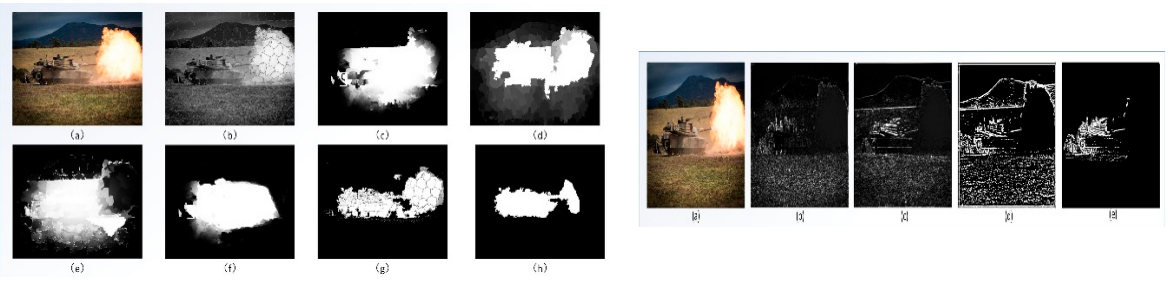
and large number, it has become one of the most preferable datasets. Most of the images in this dataset have only one significant object, so it gradually becomes the standard data set for evaluating the ability of the algorithm to handle simple scenes. ECSSD contains 1000 semantically meaningful but complex natural images. HKU-IS is another large dataset containing more than 4000 challenging images, most of which have low contrast with more than one salient object. PASCALS contains 850 challenging images (each contains multiple objects), all of which are selected from the validation set of the PASCAL VOC 2010 split dataset. The KITTI dataset is the current computer vision algorithm evaluation dataset under the largest autopilot scenario in the world. We use the first image set in the KITTI dataset "Download left color images of object data set" and the annotation file "Download training labels of object data set", most of which have multiple salient objects.

5.2. Experimental Design

First, the function of each module in the method of generating significant graphs is illustrated by algorithm performance analysis. Then, subjective and objective evaluations are conducted to highlight the superiority of the new method in comparison with the current popular saliency map generation method. Finally, we perform military object detection by using this new method, combined with efficient sub-window search method, and then compare the different detection accuracy and real-time performance among various detection methods.

5.3. Analysis of the Algorithm Performance of Object Saliency Map

The (1) in Figure 8 is the result of the saliency map generation in each stage of pre-attention channel. In the figure, (a) is the original image, (b) is the super pixel result obtained by SLIC method segmentation, and we only use the neighborhood contrast with the enhanced Sigma feature to test the saliency of (b), which generates the result (c). Result of the saliency test on the contrast of the image background is shown in (d); the spatial distance weighted saliency map results are shown in (e); the results of the local saliency map when region size is further weighted are shown in (f); the results of the global saliency map are (g), and (h) is the fusion result of the local saliency map and the global saliency map. The saliency of the tank object is prominent and the background clutter is well suppressed. Based on the above analyses, the effectiveness of the pre-attention channel for the saliency detection of military objects is verified.



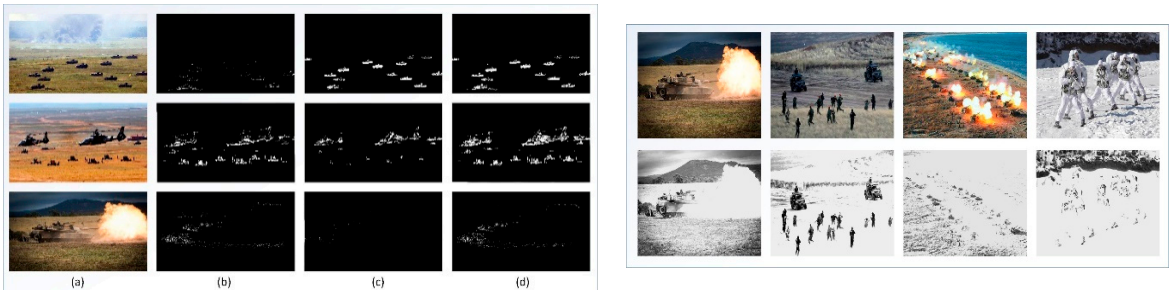
(1) **(2)**

Figure 8. The (**1**) is the result of the saliency map generation in each stage of pre-attention channel; The (**2**) is the effect of saliency detection using only texture features.

For the sub-attention channel, we verify the validity of the texture, color, and brightness features for the salient object detection, and we then verify the enhancement of the feature fusion to the significant object detection. The (2) in Figure 8, (b–d) are the results of the full-texture saliency detection of the three convolution kernels of vertical, horizontal, and circular, respectively. It can be seen that the convolution kernel that we designed is valid. The saliency of the corresponding texture is detected, but the clutter interference that is generated by the surrounding environment is also very strong. In the figure, (e) is the result of the detection of the detection area by using the detection result

of the pre-attention channel, and the pre-attention channel guidance is verified. The effectiveness of the strategy pair, and (e) is the fusion result of the three convolution kernel texture saliency detection. The texture of the tank object is prominent, and the significant interference of the noise such as flame is effectively suppressed, but some in the environment still make themselves felt strongly.

The (1) in Figure 9 is an effect diagram of saliency detection while using only color features: (b) is a red and green channel color feature saliency map, (c) is a yellow-blue channel color feature saliency map, and (d) is a fused color feature saliency map. In the first line, the red and green channel color features are not conducive to the object's saliency detection, but the yellow-blue channel color features have a better saliency detection effect. Conversely, in the second row, the yellow-blue channel color features are unfavorable for the object detection, while the red-green channel color features have a good saliency detection effect, and so the fusion features of the two have good detection results, verifying the effectiveness of color feature saliency detection and integration strategy. In addition, from the detection results of the third line of images, we found that, for camouflage objects with small differences in color characteristics and environment, tests relying only on color features cannot achieve satisfactory detection results.



(**1**) (**2**)

Figure 9. The (**1**) is the effect of saliency detection using only color features; The (**2**) is the effect of saliency detection using only brightness features.

The (2) in Figure 9 is a graph showing the effect of detecting the brightness characteristic saliency, which has a better detection effect on the military object with much difference in ambient brightness (second column), and has a better distinguishing effect on a high-brightness dominant interference object, such as a flame (the first column and the third column), but the saliency detection effect on the camouflage object is still not ideal (fourth column) and the ambient brightness clutter interference is strong.

The (1) in Figure 10 is a sub-attention channel multi-feature fusion saliency detection effect diagram, where (b) is a texture feature saliency map, (c) is a color feature saliency map, (d) is a luminance feature saliency map, and (e) is a multi-feature fusion The salient map, which verifies the validity of multi-feature fusion, can suppress interference and highlight significant objects. The (2) in Figure 10 is a pre-attention channel saliency diagram and a post-attention channel saliency diagram fusion effect diagram, where (b) is a pre-attention channel saliency diagram, (c) is a post-attention channel saliency diagram, and (d) is a two-channel fused saliency diagram, As can be seen from the figure, the object in (d) has a better detection effect.

(**1**) (**2**)

Figure 10. The (**1**) is the sub-attention channel multi-feature fusion saliency detection effect diagram; The (**2**) is the double channel fusion saliency detection effect diagram.

Figure 11 is a rendering of the significant region fusion of Gestalt visual psychology, (b), and (c) an object saliency map that is generated after the regional fusion strategy is processed, and a good detection is obtained. The effect is in line with the human visual cognition mechanism.

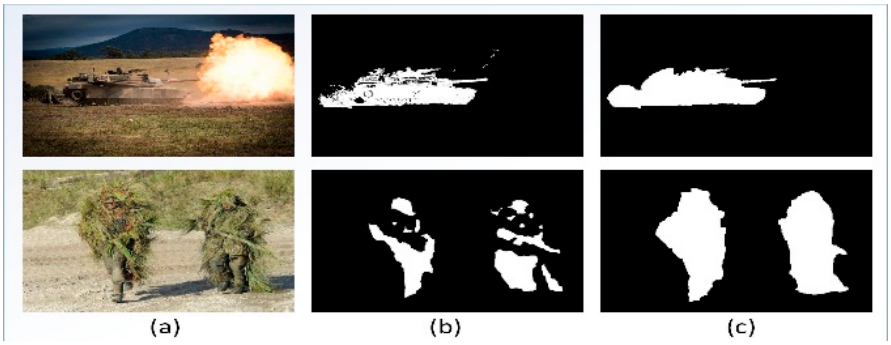


Figure 11. The rendering of the significant region fusion of Gestalt visual psychology. (**a**) Original image; (**b**) a salient region map generated by the two-channel fusion; (**c**) an object saliency map that is generated after the regional fusion strategy is processed.

5.4. Comparative Analysis of Significant Graph Generation Algorithms

In order to evaluate the superiority of the saliency map detection method on a broader context, the algorithm and 11 saliency map generation methods (RC [15], CHM [16], DSR [17], DRFI [3], MC [46], ELD [47], MDF [42], RFCN, DHS [48], DCL [49], and DSSOD [18]) are employed to compare and analyze their different significant image detection effects on various kinds of military object images. The above-mentioned algorithms are all derived from the official website open source code, while using default parameter settings.

5.4.1. Evaluation Index

We use three widely-recognized evaluation indicators to quantitatively evaluate the detection performance of the algorithm, namely the accuracy regression curve, F-measure, and mean absolute error (MAE) [18]. For a given saliency map S, we convert it to a binary map B by adaptive thresholding. The accuracy and regression rate are calculated according to the equation (37), where T represents the true detection result, P represents the precision, and R represents the regression value (Recall).

$$\begin{cases} p = \frac{|B \cap T|}{|B|} \\ R = \frac{|B \cap T|}{|T|} \end{cases} \tag{37}$$

In the equation, $|\cdot|$ represents the statistics of all non-zero inputs. The PR curve of the current data set can be obtained for the given data set and the mean of the regression rate. In order to fully assess the quality of the saliency map, the F-measure is defined as Equation (38).

$$F_\beta = \frac{\left(1 + \beta^2\right) \cdot P \times R}{\beta^2 \cdot P + R} \tag{38}$$

Here, β^2 is the weight and $\beta^2 = 0.3$ is used to emphasize the importance of the precision value. By normalizing S and T to [0, 1], we get $\hat{S}$ and $\hat{T}$, and the mean absolute error (MAE) can be defined to be Equation (39).

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| \hat{S}(i,j) - \hat{T}(i,j) \right| \tag{39}$$

Here, H and W are the height and width of the image.

5.4.2. Comparison of Visual Effects

In order to highlight the superiority and advantage of the algorithm, multiple representative images from the constructed MOD data set are selected for the comparison of saliency detection results. These images involve various kinds of environments that are difficult to detect, and they include multiple significant objects in complex and simple scenarios, significant objects of central deviation, significant objects of different sizes, and camouflage objects with low-contrasted backgrounds. We divide the selected images into several groups, which are separated by solid lines. We also provide each group with multiple labels describing its attributes. With all possible scenes taken into consideration and through a comparison of the visual effects of each algorithm, we verify that the proposed algorithm can not only highlight the correct object area, but also produce a coherent boundary in the object area, which is more in line with the visual and cognitive mechanism of the human eye. It is also worth mentioning that, due to the adaptive image enhancement based on human visual fusion and due to the saliency region detection based on dual channel and feature fusion, the significant object area saliency is enhanced, thus highlighting object and environmental background, as well as producing higher contrast. More importantly, it creates connected areas that greatly enhance the detection and expression capabilities of our model. These advantages make our results very close to the basic facts and have better detection results than other methods in almost all scenarios. The visual effect is shown in the Figure 12.

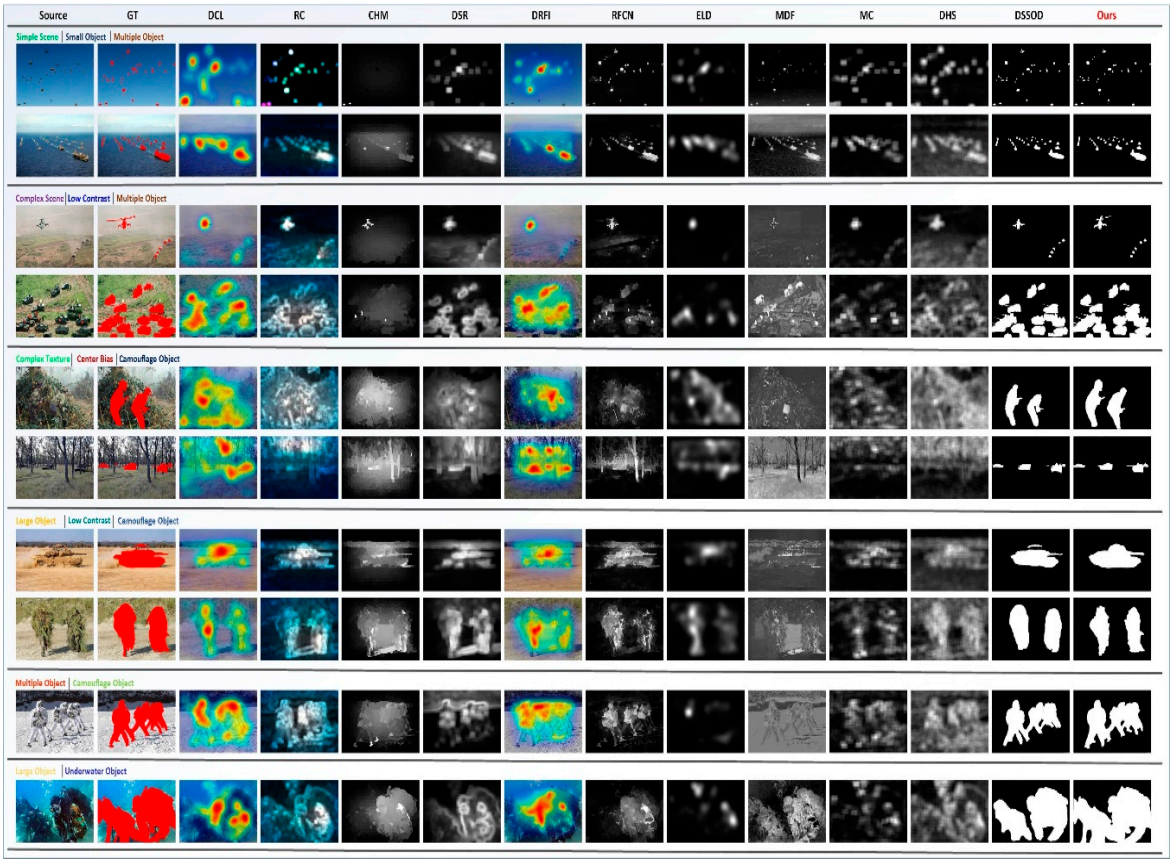


Figure 12. Selected results from various datasets. We split the selected images into multiple groups, which are separated by solid lines. To better show the capability of processing different scenes for each approach, we highlight the features of images in each group.

5.4.3. PR Curve

Here, we compare the method that was proposed with the existing ones through the PR curve. In Figure 13, we depict the PR curves that were separately generated by the methods proposed in

this paper and the most advanced methods on several popular datasets, namely, the HKU-IS, KITTI, PASCALS, and the self-contained MOD dataset. Obviously, the FCN-based method is much better than others, but the results of combining multiple data sets show that the algorithm proposed by this paper can achieve the best results. We can also find that when the recall score is close to 1, our method is much more accurate, which reflects that our false positives are much lower than other methods. This also demonstrates the effectiveness of our adaptive image enhancement, which is based on human visual fusion, as well as the effectiveness of the saliency region detection strategy based on two-channel and feature fusion, and such effectiveness renders the resulting saliency map to be closer to the basic facts.

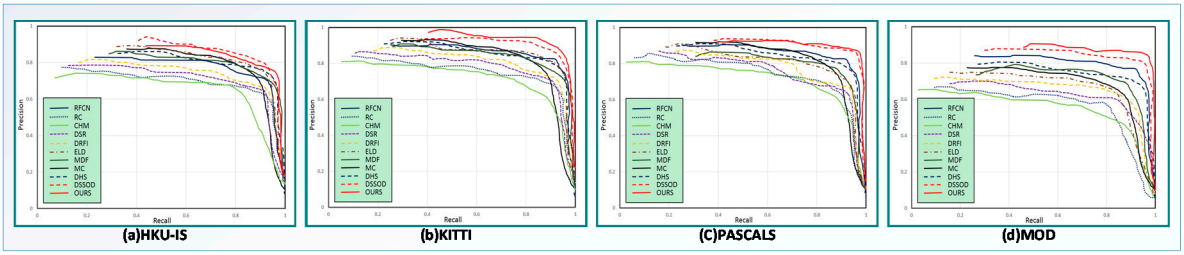


Figure 13. Precision (vertical axis) & Recall (horizontal axis) curves on three popular datasets and military object detection dataset (MOD) dataset.

5.4.4. F-Measure & MAE

We compared the method that was proposed by this paper with the existing methods in terms of F-measure and MAE scores, and the quantitative results are shown in the Table 1. For the F-measure score, the best maximum F-measure value is increased by 2% to 5% compared to other methods, which is a large margin because these values are already very close to the ideal value of 1, and the best results were obtained on MOD and KITTI datasets, both of which failed to achieve better detection results. It also verified the effectiveness and application prospects of the proposed algorithm. For the MAE score, our method achieved a reduction of more than 10% on the MOD dataset, and on other datasets there were still more than 2% reduction, which means that our number of mis-predicted cases is significantly lower than other methods. In addition, we observed that the proposed method has achieved good results on all data sets, and we verified that the proposed algorithm has better adaptive ability.

Table 1. Quantitative comparisons with 11 methods on five datasets.

Methods	PASCALS		HKU-IS		KITTI		MSRA-B		MOD	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
RFCN	0.862	0.127	0.879	0.081	0.816	0.235	0.938	0.079	0.825	0.229
RC	0.631	0.245	0.716	0.185	0.698	0.276	0.803	0.159	0.612	0.375
CHM	0.611	0.275	0.696	0.205	0.678	0.289	0.793	0.168	0.601	0.383
DSR	0.627	0.255	0.703	0.235	0.692	0.279	0.801	0.16	0.608	0.379
DRFI	0.812	0.149	0.819	0.131	0.801	0.255	0.938	0.079	0.725	0.279
ELD	0.768	0.121	0.843	0.073	0.816	0.168	0.913	0.041	0.795	0.208
MDF	0.765	0.147	0.859	0.131	0.798	0.169	0.884	0.115	0.763	0.265
MC	0.723	0.149	0.792	0.103	0.732	0.185	0.872	0.063	0.691	0.325
DHS	0.819	0.101	0.892	0.052	0.785	0.201	0.906	0.058	0.739	0.298
DSSOD	0.83	0.08	0.913	0.039	0.896	0.045	0.927	0.028	0.803	0.196
OURS	0.852	0.061	0.896	0.041	0.912	0.031	0.915	0.068	0.875	0.082

5.5. Comparative Analysis of Algorithm Detection Results

There are objects $Z\{z_1, z_2, \cdots, z_n\}$ in the image, where $z_i = \left[x_z^i, y_z^i, a_z^i, b_z^i, c_z^i, d_z^i\right]$. It is assumed that the algorithm outputs the image is $W\{w_1, w_2, \cdots, w_m\}$, and $w_j = \left[x_w^j, y_w^j, a_w^j, b_w^j, c_w^j, d_w^j\right]$, x, y, a, b

of the object are separately the coordinates of the upper left corner point, the width and height of the object's bounding box. c is the object's confidence, and d is the category to which the object belongs. If so, the evaluation process includes the following steps:

(1) Establish an optimal one-to-one correspondence between goals and hypothetical results. The Euclidean distance is used to calculate the spatial position correspondence between the real object and the hypothetical object. The threshold T of the Euclidean distance is set to the distance between the centers of the hypothesis and the least overlap of the objects. The number of objects for completing the correspondence is NT, and the number of missed objects is $LP = n - NT$.

(2) After the mutual correspondence between the objects is completed, we divide the test results into two categories according to d, which is in accordance with the real object and the hypothetical object. The two categories are: accurate detection, which occurs when there are same categories; and, misdetection, which occurs when there are different categories. The number of objects accurately detected by statistics is TR, and the number of objects for statistical misdetection is TW. When we compare the number of real objects $\boldsymbol{n}$ with the number of objects detected m, and if $n < m$, there is a case of false alarms, and the number of false objects is $FP = m - NT$.

(3) From the statistical results of step 2, we can measure the detection effect of the algorithm by calculating the false alarm rate, missed alarm rate, detection rate, and false detection rate of the algorithm.

False alarm rate, $P_f = \frac{FP}{n}$; Missed alarm rate, $P_m = \frac{LP}{n}$;
Detection rate, $P_d = \frac{TR}{n}$; false detection rate, $P_e = \frac{TW}{n}$.

Deep learning is a learning method based on deep artificial neural network. In various types of artificial neural network structures, deep convolutional networks have powerful feature extraction capabilities. In visual tasks, such as image recognition, image segmentation, object detection, and scene classification, very good results have been achieved, so this paper selects Faster R-CNN [50], DSOD300 [51] detection framework, both of which are based on deep neural network, as well as YOLOv2 544 [52] and model DSSD [53].

Faster R-CNN (where R corresponds to "Region") is the best method based on deep learning R-CNN series object detection. Training includes four steps: (1) Initialize the RPN network parameters using a pre-trained model on ImageNet to fine tune the RPN network; (2) Using the RPN network in (1) to extract the region proposal to train the Fast R-CNN network, and also initialize the network parameters with the pre-trained model on ImageNet; (now it seems that the two networks are relatively independent); (3) Re-initialize the RPN using the Fast R-CNN network of (2), fine-tune the fixed convolutional layer, and fine-tune the RPN network; and, (4) Fix the convolution layer of Fast R-CNN in (2), and fine-tune the Fast R-CNN network by using the region proposal extracted by RPN in (3).

The DSOD network structure can be divided into two parts: the backbone sub-network for feature extraction and the front-end sub-network for prediction over multi-scale response maps. The DSOD model not only has fewer parameters and better performance, but it also does not need to be pre-trained on large data sets (such as ImageNet), which makes the design of the DSOD network structure flexible and make it possible to design its own network structure according to its own application scenario. The training parameters are roughly, as follows: learning rate = 0.1, decreasing to 0.1 every 20,000 ITER, snapshot every 2000 ITER, maximum ITER 100,000, and test every 2000 ITER.

Yolov2 treats the object detection task as a regression problem, and it uses a neural network to directly predict the coordinates of the bounding box, the confidence of the object contained in the box, and the probabilities of the object from a whole image. Because Yolo's object detection process is completed in a neural network, it is possible to optimize the object detection performance.

DSSD has poor robustness against small targets, and it changes SSD's reference network from VGG to ResNet-101, thus enhancing feature extraction capability. The use of a de-convolution layer provides a great deal of context information. Most of the training techniques are similar to the original SSD. First, SSD's Default Boxes are still used, and those with an overlap rate higher than 0.5 are

considered positive samples. Then we set some more negative samples so that the ratio of positive samples to negative samples is 3:1. The joint loss function of smooth L1 + soft max is minimized during training. Data expansion is still needed before training (including the hard example mining technique). In addition, the default Boxes dimension of the original SSD was manually specified and it may not be efficient enough. Therefore, seven default boxes dimensions were obtained again by K-means clustering method, and the obtained boxes dimensions are more representative.

All of the detection frameworks use the default parameter settings in the official code published by the author. The detection categories were adjusted, trained, and tested by using the KITTI and MOD data sets.

As a comparison algorithm of object detection performance, the object detection effect is verified on the MOD and KITTI data sets. The detection effect is shown in Table 2, where T_{ime} represents the average time taken by the algorithm to process an input image with a single frame size of 480 × 480 in the dataset.

Table 2. The comparisons with five methods on five datasets.

Methods	Dataset	P_f (%)	P_m (%)	P_d (%)	P_e (%)	T_{ime} (s)
Faster R-CNN	KITTI	11.21	13.34	60.32	15.13	0.076
	MOD	16.25	15.38	50.83	17.54	0.086
DSOD300	KITTI	14.48	15.91	63.38	6.23	0.017
	MOD	18.95	19.28	51.42	10.35	0.021
YOLO V2 544	KITTI	13.31	12.29	59.84	14.56	0.022
	MOD	15.17	15.49	49.45	19.89	0.026
DSSD	KITTI	9.53	10.69	66.25	13.53	0.018
	MOD	16.24	13.19	55.16	15.41	0.028
OURS	KITTI	4.19	3.13	90.47	2.21	0.106
	MOD	6.16	5.37	82.05	6.42	0.185

Here, we compare the results of our algorithm and other deep learning comparison algorithms in the above table: in the KITTI data set, the object detection rate increases by 24~32%, reaching 90.47%, and total time cost of the single frame image algorithm processing is about 0.106 s; in the WD data set, the object detection rate increases by 27~33%, reaching 82.05%, and the overall time cost of the single frame image algorithm is about 0.185 s. With two data sets combined, it is clear that the saliency region detection process that is based on dual channel and feature fusion takes about 0.1 s, and the proportion of time consumed for adaptive image enhancement process based on human visual fusion against the complexity of the image is about 0.03~0.08 s. In general, the proposed algorithm achieves a good balance between detection accuracy and real-time performance in comparison with other object detection algorithms.

5.6. Discussion

Figure 14 is a diagram showing the effect of the algorithm on the detection of military objects in various scenarios. The purple border is the underwater object detection effect map; the light blue border is the detection effect map of the large object, multiple objects, and small objects, all of which are interfered by the environment. the red border is the detection effect map of the camouflage object; the green border and the dark blue border verify the effectiveness of the adaptive overlap threshold strategy in detecting the border fusion process. However, as for the objects with close airspace distance, some object detection frames will blend with each other, which subsequently needs further research.

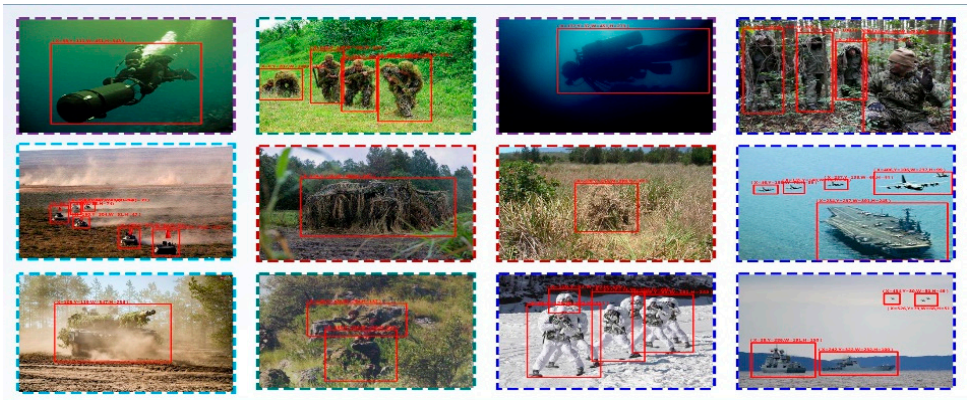


Figure 14. The diagram shows the effect of the algorithm on the detection of military objects in various scenarios.

6. Conclusions

This paper proposes a military object detection method that combines human visual saliency and visual psychology. Firstly, the adaptive adjustment mechanism of the human eye is modeled, and a new image adaptive enhancement method based on human visual fusion is proposed, which can effectively highlight the object and suppress the interference. Then, inspired by how the human visual information is processed, we establish the distinctive region detection model that is based on dual channel and feature fusion. After that, the pre-attention channel and the post-processing channel, respectively, perform significant region detection on the image, and after the two channel results get fused, the measurement result determines the candidate salient region. Later on, under the guidance of Gestalt visual psychology, we integrate the obtained candidate salient regions to obtain an object saliency map with overall perception, and then apply the efficient sub-window search method to detect and screen the object so as to identify the object location and region range. Experiments show that our algorithm can realize the rapid and accurate detection of military objects in various complex scenes, reduce effectively the camouflage and deception effect of the battlefield, create favorable conditions for implementing precision strikes, and have great prospects for future application.

Author Contributions: Conceptualization, X.H. (Xia Hua) and X.W.; Methodology, X.H. (Xia Hua) and X.W.; Software, X.H. (Xia Hua); Validation, X.H. (Xia Hua), X.W., D.W.; Formal Analysis, J.H.; Investigation, X.H. (Xia Hua), J.H. and X.H. (Xiaodong Hu); Resources, X.W., D.W.; Data Curation, X.H. (Xia Hua); Writing-Original Draft Preparation, X.H. (Xia Hua), X.W., J.H. and X. H. (Xiaodong Hu); Writing-Review & Editing, X.W.; Visualization, X.H. (Xia Hua); Supervision, X.H. (Xia Hua); Project Administration, X.W.; Funding Acquisition, X.W.

Funding: This work was supported in part by the China National Key Research and Development Program (No. 2016YFC0802904), National Natural Science Foundation of China (61671470), Natural Science Foundation of Jiangsu Province (BK20161470), 62nd batch of funded projects of China Postdoctoral Science Foundation (No. 2017M623423).

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

References

1. Sun, Y.; Chang, T.; Wang, Q.; Kong, D.; Dai, W. A method for image detection of tank armor objects based on hierarchical multi-scale convolution feature extraction. *J. Ordnance Eng.* **2017**, *38*, 1681–1691.
2. Dollar, P.; Appel, R.; Belongie, S.; Perona, P. Fast feature pyramids for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 32–45. [CrossRef] [PubMed]
3. Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S. Salient object detection: a discriminative regional feature integration approach. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2083–2090.

4. Schneiderman, H. Feature-centric evaluation for efficient cascaded object detection. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), Washington, DC, USA, 27 June–2 July 2004; Volume 2, pp. II-29–II-36.
5. Li, L.; Huang, W.; Gu, I.Y.-H.; Tian, Q. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans. Image Process.* **2004**, *13*, 1459–1472. [CrossRef] [PubMed]
6. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video Processing from electro-optical sensors for object detection and tracking in a maritime environment: A Survey. *IEEE Trans. Intell. Trans. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
7. Savaş, M.F.; Demirel, H.; Erkal, B. Moving object detection using an adaptive background subtraction method based on block-based structure in dynamic scene. *Optik* **2018**, *168*, 605–618. [CrossRef]
8. Sultani, W.; Mokhtari, S.; Yun, H.B. Automatic pavement object detection using superpixel segmentation combined with conditional random field. *IEEE Trans. Intell. Trans. Syst.* **2018**, *19*, 2076–2085. [CrossRef]
9. Zhang, C.; Xie, Y.; Liu, D.; Wang, L. Fast threshold image segmentation based on 2D fuzzy fisher and random local optimized QPSO. *IEEE Trans. Image Process.* **2017**, *26*, 1355–1362. [CrossRef] [PubMed]
10. Druzhkov, P.N.; Kustikova, V.D. A survey of deep learning methods and software tools for image classification and object detection. *Pattern Recognit. Image Anal.* **2016**, *26*, 9–15. [CrossRef]
11. Ghesu, F.C.; Krubasik, E.; Georgescu, B.; Singh, V.; Zheng, Y.; Hornegger, J.; Comaniciu, D. Marginal space deep learning: efficient architecture for volumetric image parsing. *IEEE Trans. Med. Imaging* **2016**, *35*, 1217–1228. [CrossRef] [PubMed]
12. Xu, X.; Li, Y.; Wu, G.; Luo, J. Multi-modal deep feature learning for RGB-D object detection. *Pattern Recognit.* **2017**, *72*, 300–313. [CrossRef]
13. Schölkopf, B.; Platt, J.; Hofmann, T. Graph-based visual saliency. In Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; MIT Press: Cambridge, MA, USA, 2006; pp. 545–552.
14. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–26 June 2009; pp. 1597–1604.
15. Cheng, M.M.; Mitra, N.J.; Huang, X.; Torr, P.H.S.; Hu, S.M. Global contrast based salient region detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 569–582. [CrossRef] [PubMed]
16. Li, X.; Li, Y.; Shen, C.; Dick, A.; Hengel, A.V.D. Contextual hypergraph modeling for salient object detection. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 3328–3335.
17. Li, X.; Lu, H.; Zhang, L.; Ruan, X.; Yang, M.H. Saliency detection via dense and sparse reconstruction. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2976–2983.
18. Hou, Q.; Cheng, M.M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P.H.S. Deeply supervised salient object detection with short connections. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**. [CrossRef] [PubMed]
19. Wang, L.; Wang, L.; Lu, H.; Zhang, P.; Ruan, X. Saliency detection with recurrent fully convolutional networks. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; Volume 9908.
20. Dresp, B.; Grossberg, S. Contour integration across polarities and spatial gaps: From local contrast filtering to global grouping. *Vis. Res.* **1997**, *37*, 913–924. [CrossRef]
21. Dresp, B.; Durand, S.; Grossberg, S. Depth perception from pairs of stimuli with overlapping cues in 2-D displays. *Spat. Vis.* **2002**, *15*, 255–276. [CrossRef] [PubMed]
22. Dresp-Langley, B.; Grossberg, S. Neural computation of surface border ownership and relative surface depth from ambiguous contrast inputs. *Front. Psychol.* **2016**, *7*, 1102. [CrossRef] [PubMed]
23. Grillspector, K.; Malach, R. The human visual cortex. *Ann. Rev. Neurosci.* **2004**, *27*, 649–677. [CrossRef] [PubMed]
24. Blog. Available online: https://blog.csdn.net/shuzfan/article/details/78586307 (accessed on 6 August 2018).
25. Wagemans, J.; Feldman, J.; Gepshtein, S.; Kimchi, R.; Pomerantz, J.R.; van der Helm, P.A.; van Leeuwen, C. A century of gestalt psychology in visual perception II. conceptual and theoretical foundations. *Psychol. Bull.* **2012**, *138*, 1218–1252. [CrossRef] [PubMed]

26. Lee, T.S. Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 959–971.
27. Zhihu. Available online: https://zhuanlan.zhihu.com/p/21905116 (accessed on 6 August 2018).
28. Stocker, A.A.; Simoncelli, E.P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **2006**, *9*, 578–585. [CrossRef] [PubMed]
29. Kastner, S.; Pinsk, M.A. Visual attention as a multilevel selection process. *Cognit. Affect. Behav. Neurosci.* **2004**, *4*, 483–500. [CrossRef]
30. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
31. Liu, S.T.; Liu, Z.X.; Jiang, N. Object segmentation of infrared image based on fused saliency map and efficient subwindow search. *Acta Autom. Sin.* **2018**, *11*, 274–282.
32. Lan, Z.; Lin, M.; Li, X.; Hauptmann, A.G.; Raj, B. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 204–212.
33. Cacioppo, J.T.; Petty, R.E.; Kao, C.F.; Rodriguez, R. Central and peripheral routes to persuasion: An individual difference perspective. *J. Person. Soc. Psychol.* **1986**, *51*, 1032–1043. [CrossRef]
34. Tuzel, O.; Porikli, F.; Meer, P. Region Covariance: A fast descriptor for detection and classification. In *Computer Vision—ECCV 2006*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 589–600.
35. Cela-Conde, C.J.; Marty, G.; Maestú, F.; Ortiz, T.; Munar, E.; Fernández, A.; Roca, M.; Rosselló, J.; Quesney, F. Activation of the prefrontal cortex in the human visual aesthetic perception. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 6321–6325. [CrossRef] [PubMed]
36. Liang, D. Research on Human Eye Optical System and Visual Attention Mechanism. Ph.D. Thesis, Zhejiang University, Hangzhou, China, 2017.
37. Hong, X.; Chang, H.; Shan, S.; Chen, X.; Gao, W. Sigma Set: A small second order statistical region descriptor. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–26 June 2009; pp. 1802–1809.
38. Lauinger, N. The two axes of the human eye and inversion of the retinal layers: The basis for the interpretation of the retina as a phase grating optical, cellular 3D chip. *J. Biol. Phys.* **1993**, *19*, 243–257. [CrossRef]
39. Dong, L.; Wesseloo, J.; Potvin, Y.; Li, X. Discrimination of mine seismic events and blasts using the fisher classifier, naive bayesian classifier and logistic regression. *Rock Mech. Rock Eng.* **2016**, *49*, 183–211. [CrossRef]
40. Fang, Z.; Cui, R.; Jin, W. Video saliency detection algorithm based on bio-visual features and visual psychology. *Acta Phys. Sin.* **2017**, *66*, 319–332.
41. Liu, T.; Sun, J.; Zheng, N.-N.; Tang, X.; Shum, H.-Y. Learning to detect a salient object. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 353–367. [PubMed]
42. Li, G.; Yu, Y. Visual saliency based on multiscale deep features. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5455–5463.
43. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
44. Li, Y.; Hou, X.; Koch, C.; Rehg, J.M.; Yuille, A.L. The secrets of salient object segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 280–287.
45. Che, Z.; Zhai, G.; Min, X. A hierarchical saliency detection approach for bokeh images. In Proceedings of the 2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP), Xiamen, China, 19–21 October 2015; pp. 1–6.
46. Zhao, R.; Ouyang, W.; Li, H.; Wang, X. Saliency detection by multi-context deep learning. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1265–1274.
47. Lee, G.; Tai, Y.W.; Kim, J. Deep saliency with encoded low level distance map and high level features. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 660–668.

48. Liu, N.; Han, J. DHSNet: Deep hierarchical saliency network for salient object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 678–686.
49. Li, G.; Yu, Y. Deep contrast learning for salient object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 478–487.
50. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
51. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.G.; Chen, Y.; Xue, X. DSOD: Learning deeply supervised object detectors from Scratch. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1937–1945.
52. Zhang, J.; Huang, M.; Jin, X.; Li, X. A real-time chinese traffic sign detection algorithm based on modified YOLOv2. *Algorithms* **2017**, *10*, 127. [CrossRef]
53. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. SSD: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.