Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 2412 – 2429
© TÜBİTAK
doi:10.3906/elk-1808-208

Spatial-aware global contrast representation for saliency detection

Dan XU*, **Shucheng HUANG**, **Xin ZUO**
School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, China

Received: 30.08.2018 • **Accepted/Published Online:** 06.03.2019 • **Final Version:** 26.07.2019

Abstract: Deep learning networks have been demonstrated to be helpful when used in salient object detection and achieved superior performance than the methods that are based on low-level hand-crafted features. In this paper, we propose a novel spatial-aware contrast cube-based convolution neural network (CNN) which can further improve the detection performance. From this cube data structure, the contrast of the superpixel is extracted. Meanwhile, the spatial information is preserved during the transformation. The proposed method has two advantages compared to the existing deep learning-based saliency methods. First, instead of feeding the deep learning networks with raw image patches or pixels, we explore the spatial-aware contrast cubes of superpixels as training samples of CNN. Our method is superior because the saliency of a region is more dependent on its contrast with the other regions than its appearance. Second, to adapt to the diversity of a real scene, both the color and textural cues are considered. Two CNNs, color CNN and textural CNN, are constructed to extract corresponding features. The saliency maps generated by two cues are concatenated in a dynamic way to achieve optimum results. The proposed method achieves the maximum precision of 0.9856, 0.9250, and 0.8949 on three benchmark datasets, MSRA1000, ECSSD, and PASCAL-S, respectively, which shows an improvement of performance in comparison to the state-of-the-art saliency detection methods.

Key words: Saliency detection, convolutional neural networks, spatial-aware, global contrast cube

1. Introduction

Human beings have the ability to focus on the prominent parts of an image without great effort. Saliency detection research is trying to model this ability by studying the human vision mechanism and psychology [1, 2]. The topic has drawn great attention in computer vision for its wide range of application areas in object detection [3], image classification [4], image retrieval [5], image/video compression and retargeting [6]. The processing of the salient region detection is helpful for saving memory space and computation costs.

Saliency detection differs from the related studies on human eye-fixation prediction and object proposal. The human eye-fixation prediction aims at predicting the gaze or movement of human eyes, which is often used for intelligent human-computer interactions[7]. Object proposal mainly focuses on the evaluation of the objective probabilities of the given bounding boxes. It tends to highlight all the possible bounding boxes instead of the exact salient objects[8].

The traditional bottom-up saliency detection methods are usually based on the low-level image features and prior knowledge[9]. The center-bias prior is based on the hypothesis that the salient objects are often focused on the center of an image instead of the edges. Spatial compactness is derived from the fact that the background usually occupies a large range of the image and the salient objects tend to have compact spatial

*Correspondence: xudan_zj@163.com

extensions. Contrast is based on the human vision mechanism through which people tend to be attracted to the objects with high contrast against the surroundings. This prior knowledge and low-level hand-crafted features are applied in heuristic ways, which work well for most sample context images and with expected performance. However, they are not very helpful for highlighting objects from the complex scene in real-world images.

In this paper, we utilize a convolution neural network (CNN) to learn the high-level features from the low-level features and presumptive assumptions in a data-driven way, with the benefit of obtaining hierarchical features automatically as the human brain does. Instead of raw image patches, we use the global contrast of superpixels as the training samples of our deep learning networks. The reason is that we define a region as salient or not by its contrast with other regions rather than its own characteristics. Visual attention studies show that the spatial distance with the context mainly influences the saliency of an object. Motivated by these studies, we introduce a spatial-aware contrast cube in which the first two dimensions represent the original spatial position of the superpixels and the third dimension denotes their feature dissimilarity.

The main contributions of this paper are listed as follows. 1) We propose a spatial-aware cube representation of the superpixels, which are considered to be the training samples of the deep learning networks, with the advantage of extracting global contrast features and preserving the spatial information as well. 2) We take color and textural cues into account simultaneously. Color cues play an important role in saliency detection; however, they may fail in the texture-dominant images. To address this problem, we propose the texture-based CNN to capture the textural features and compensate for the drawback of the color contrast method. The saliency maps that are produced by color and textural cues are integrated in a dynamic way to obtain optimal performance.

This paper is organized as follows. In Section 2 the related work and present researches are given. In Section 3 the proposed method is described in detail with subsections of spatial-aware contrast cube, color contrast CNN, textural contrast CNN, and saliency fusion. Experiments and discussions are given in Section 4. In the last section, the conclusions and future work are discussed.

2. Related work

Visual saliency was derived from the research of neurobiology and then introduced to the tasks of computer vision [10]. Over the past two decades, a variety of algorithms have been proposed to compute the visual saliency from digital images and videos. The development of the saliency detection study can be roughly divided into four stages. The representative work in the first stage was the pioneering research of Itti et al., which realized the visual biological model using the bottom-up computational method [11]. The second stage began in 2007 and was represented by the work of Liu et al. [12]. It was the first time that saliency detection was considered a binary segmentation problem. In this stage, the saliency detection eliminated the biological model and tended toward the low-level feature-based computational model, which enhanced the performance dramatically [12–14]. The third stage was featured by the fusion of various saliency approaches. The task-driven top-down approaches introduced high semantic information such as the human face and warm colors, and combined with the preattentive bottom-up models, they were able to improve the performance to some extent. With the same purpose, diverse saliency cues, including uniqueness, compactness, and objectness, were fused to achieve better precision and recall results [9, 15–17]. In the last stage, deep learning was introduced to saliency detection to extract high-level features and alleviate the dependency on the hand-crafted low-level features [18, 19].

Global contrast can help highlight the uniform salient regions from the background and generate full-size saliency maps. The histogram contrast (HC) method utilizes the global statistic features of the image to evaluate the saliency value of each pixel based on the distance between the color histogram bins [2]. Tong et al. proposed a method that considers the global and local contrast in a saliency framework in which the global contrast is computed first as the initial saliency map and then a locality-constrained linear coding method is used for optimization [20]. The aforementioned contrast methods will highlight the background regions instead of the salient objects for images containing large salient objects because the color distances among regions are accumulated in an unweighted manner. To address this problem, Xu et al. improved the global contrast model by introducing a weighted mean vector, which can effectively repress the background and highlight the large salient objects in the saliency maps [21].

Previous studies have demonstrated that the spatial distance among regions mainly affects the region saliency. The region contrast (RC) method improves the HC by considering the saliency based on the regions instead of pixels to help obtain meaningful salient objects and save computation costs. Another improvement is that RC takes the spatial feature into account and holds the idea that the closer regions have the greater influence on the region saliency. All the improvements help RC to be one of the best traditional saliency detection methods. According to visual organization principles, there are some attention centers around which the visual form is structured. The salient objects tend to cluster together around the attention centers. In contrast, the background regions usually have similar visual features with both closer and farther regions. Goferman et al. defined a distance measurement to evaluate the spatial saliency of a pitch based on the principle [14]. Xu et al. introduced the concept of spatial relative spread, which computes the cluster spatial saliency by measuring the distance with other clusters in the entire image[22]. In [23], a spatial coherence model that assigns consistent saliency values for adjacent superpixels without strong edges is applied to saliency maps to achieve further performance improvements.

Deep learning-based methods aim at learning hierarchical features similar to the way in which humans organize information. The methods have been shown to be superior to the works that rely on low-level features and prior hypotheses when used for saliency detection. In [24], a multicontext deep learning model that combines the global context and local context is proposed, which is followed by a pretraining stage to further refine the saliency detection results. In [18], local contrast features are learned by a local deep neural network, which is incorporated with high-level object proposal and global contrast features to acquire the final candidate regions. Both of the above methods take the local and global contrast into account and include additional refinement for deep learning saliency results. Lee et al. consider the complementary of features at different levels instead. The low-level hand-crafted features such as the color component value, texture, and color histogram are integrated into a distance map to complement the high-level features generated by VGG-net [19].

Saliency detection is helpful for present researches and applications related to computer vision. In [25], a geodesic distance-based saliency detection method was explored to offer object-level cues for unsupervised video segmentation. In [6], saliency map was used as a component of energy function to estimate the importance of pixels, followed by producing a visual quality retargeted image. Kutbay et al. proposed a bottom-up saliency mapping with quaternion convolution to measure carotid artery intima-media thickness, which can be developed as a part of a computer-aided diagnosis system [26]. Bonnin-Pascual et al. explored a defect detection approach based on saliency-related features to detect cracks or corrosion on vessels [27].

3. The proposed method

This study aims at improving saliency detection performance with a novel spatial-aware global contrast representation and deep learning networks.The framework of the proposed method is shown in Figure 1. SLIC

[28] is utilized to segment input image into regions with advantages of adhering to boundaries and resulting in compact superpixels. Visual samples of SLIC superpixel method are given in Figure 2. For each region, both color and texture features are extracted to construct spatial-aware contrast cubes respectively. Then color and textual CNNs are built and fed with contrast cubes to extract high-level features. Softmax function is used after CNN to evaluate the probability of a region being salient or not. Saliency maps generated by color and textual cues are fused in a dynamic way to refine the results.

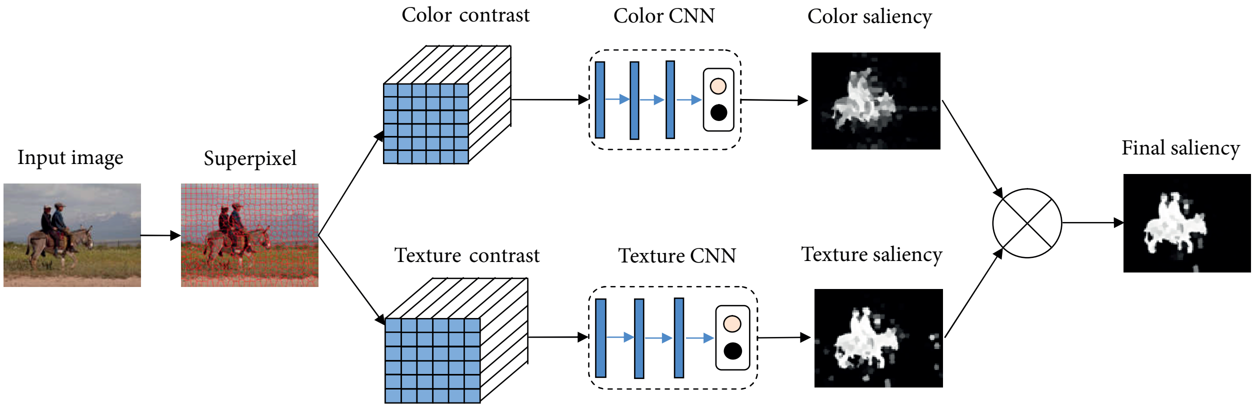


Figure 1. The framework of our method.

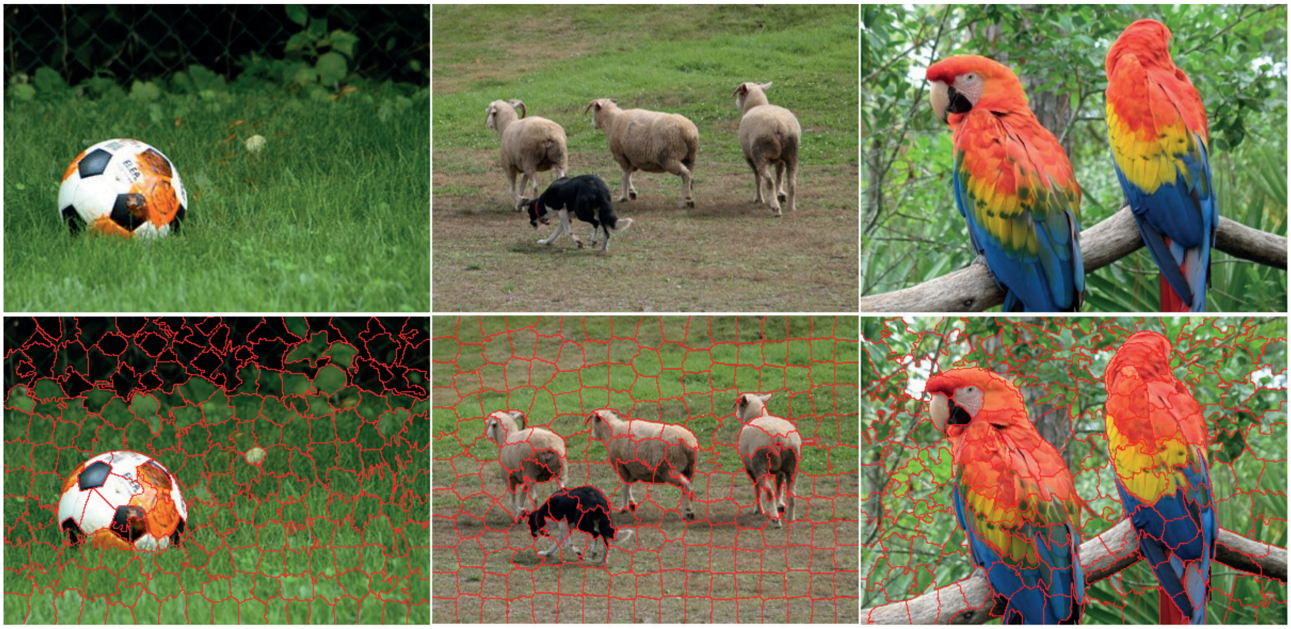


Figure 2. Visual samples of SLIC superpixel method. The first row is original images and the second row is segmented superpixels.

3.1. Spatial-aware contrast cube

When deep learning is used for saliency detection, the training sample is an important factor of the whole method since it influences the performance to a large extent. Typically, a series of raw local image patches are used as training samples to feed the deep learning network. Each training sample is of size $n \times n \times 3$, where n is the width and height of the patch and 3 represents the three channels of R, G, and B. The trained network extracts the color features of the images without consideration of other cues of the saliency, such as the texture, spatial information, and contrast.

He et al. utilized a meaningful contrast sequence as the input of the deep learning network [1]. The sequence is a novel data structure for training the network, which captures the contrast features of the superpixels effectively. However, when the 2D image is transferred to a 1D contrast sequence, the spatial information of superpixels is lost. On the other hand, the exact number of image superpixels generated by SLIC cannot be controlled. Therefore, the transformation will result in different lengths of sequences for different images.

In this paper, we propose a spatial-aware contrast cube that can preserve the 2D spatial information as the input of the deep learning network. Suppose that there are Q superpixels in an image, and we divide the image into $N \times N$ regions, where $Q \approx N \times N$. Then, we match superpixel sp_i with region r_i. r_i is represented by the superpixel with the max area included in it in case of straddling two or more superpixels. Therefore, all the images, regardless of their original size, will be set to the size of $N \times N$ without cropping or resizing. The spatial-aware contrast cube of region r_i can be obtained by computing the feature distance between its corresponding superpixel sp_i and the superpixels of the other regions in the image. The size of the contrast cube is $N \times N \times M$, where M is the length of the feature vector.

Figure 3 shows the proposed global contrast cube. Taking region r_{23} as an example, its corresponding superpixel s_7 occupies the maximum area of the region. The contrast cube of r_{23} is a three-dimensional feature distance matrix between s_7 and the other superpixels in the image. We take the uniform contrast cubes as the training samples of our CNN.

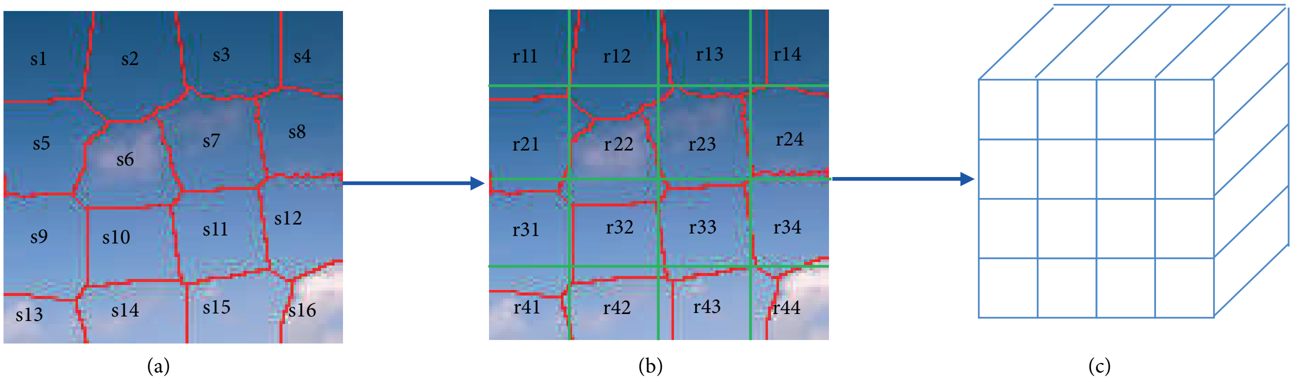


Figure 3. Spatial-aware contrast cube. From left to right are superpixel segmentation, relation between regions and the corresponding superpixels, and spatial-aware contrast cube of a region.

By the spatial-aware contrast cube, we obtain uniform training samples of size $N \times N \times M$ without resizing or rescaling the original images. In addition, the positional information of the superpixels is preserved during the calculation of global contrast, both of which are essential for saliency detection.

3.2. Color contrast CNN

The visual attention mechanism shows that the saliency is mainly derived from the uniqueness and rarity, which can be captured by measuring the contrast among regions with specific image attributes such as the color or gradient. Obviously, the feature descriptor plays an essential role during the measurement of the contrast.

When the color descriptor is concerned, it should be discriminative and insensitive to illumination. The color name is a color descriptor that includes 11 color terms that originate from the human language, which are black, blue, brown, gray, green, orange, pink, purple, red, white, and yellow [29]. Color name learning was executed to explore the relation between the pixel values and the color names in our previous work [22]. Given a pixel x with Pixel value $f(x)$ in lab color space its color name descriptor is an 11-dimensional vector where every component gives the probability of the corresponding color name. Accordingly, the color name descriptor (CN) of a superpixel sp is interpreted as a vector involving 11 color names' conditional probability as follows:

$$CN = (p(cn_1|sp), p(cn_2|sp), \cdots, p(cn_{11}|sp)) \tag{1}$$

with

$$p(cn_i|sp) = \frac{1}{R}\sum_{x\in sp} p(cn_i|f(x)), \tag{2}$$

where R is the number of pixels in sp, and $p(cn_i|f(x))$ is the conditional probability of the color term i for the given $f(x)$, which can be acquired from the color name learning.

To present the discriminability of color names, we take the image in Figure 4a as an example and evaluate the color name features of its background and foreground. Although the salient object (the leopard) has a similar color with the background, the color name features capture the differences between them, where the background is featured by green and brown as shown in Figure 4b while the object is predominatly black and brown as shown in Figure 4c.

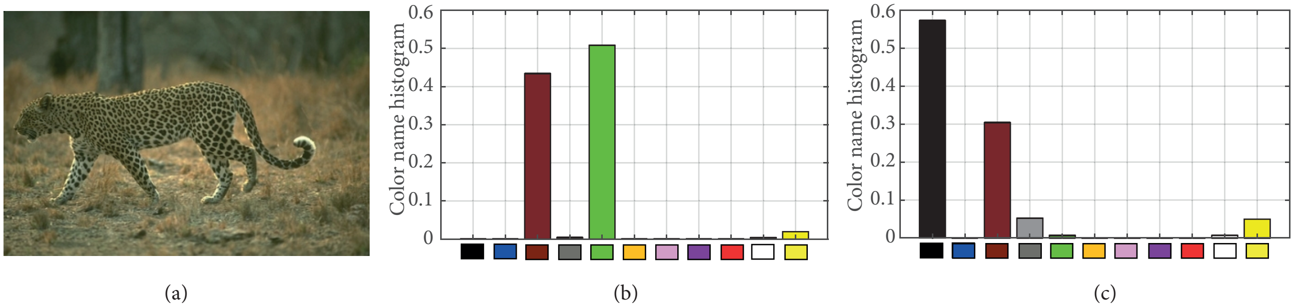


Figure 4. Color name features of an image. (a) Original image, (b) color name distributions of the background, and (c) color name distributions of the salient object.

The closer superpixels may have greater influences on the saliency values of a superpixel than the farther ones. We emphasize such a spatial layout of information by adding the superpixel average coordination c_x and c_y to the feature list to construct the 13-dimensional feature vector. There are 24×24 regions in an image totally; therefore, the size of the contrast cube of a region is $24 \times 24 \times 13$.

Our CNN structure consists of nine layers that can be assigned to four stages. The first stage includes the convolution, max pooling and ReLu nonlinearity function. The second stage is exactly the same as the first one except for the dimension of the convolution layer. The third stage contains the convolution and ReLu nonlinearity functions, while the last stage only contains a filter bank layer. The whole CNN is followed by a softmax layer to evaluate the probability of a superpixel being salient or not. The convolution layer extracts the features among local regions. More convolution layers will acquire features in larger ranges. The max pooling layer is used for feature convergence, dimension reduction, and slight translation invariance. The

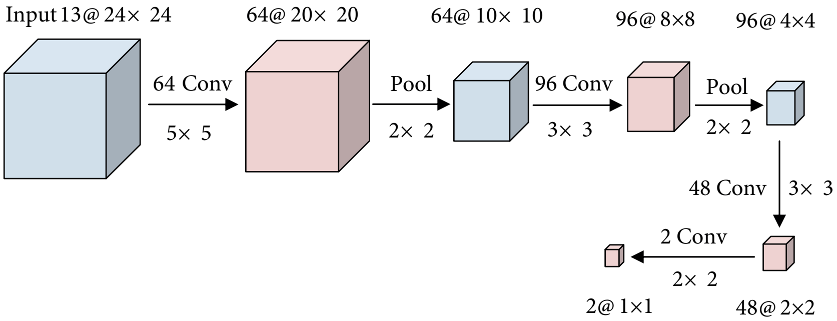


Figure 5. The structure of the color contrast CNN.

ReLu nonlinearity function introduces the nonlinear features into the network. It is also helpful to reduce the computational complexity and memory space with its sparsity. Note that the nonlinearity function does not change the dimensions of the training data; therefore, we ignore the ReLu layers in Figure 5 for display reasons.

The CNN is fed with a $24 \times 24 \times 13$ spatial-aware contrast cube for a superpixel, in which 24×24 represents the number of regions in an image and 13 denotes the feature dimensions.

Given a training superpixel set S_i and the corresponding label set l_i, we use the cross-entropy with weight decay as the loss function:

$$L(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=0}^{1} 1(l_i = j) log P(l_i = j|\theta) + \lambda \sum_{k=1}^{K} \|W_k\|_F^2, \tag{3}$$

where θ is the parameter set of our network, which includes the weight and bias vectors of all the convolution layers. j denotes the label of 0 or 1, and $1\{l_i = j\}$ is an indicator function that gives the ground truth. $P(l_i = j|\theta)$ is the conditional probability of the ith superpixel being salient as generated by the softmax loss. λ is the weight decay factor with value of 0.0005 and K is the number of convolution layers of CNN with value of 4 in our implementation. W_k is the weight of the kth convolution layer. $\sum_{k=1}^{K} \|W_k\|_F^2$ represents the L2-regularization of W_k. We use the stochastic gradient descent (SGD) method to train our network with a batch size of $m = 256$ and the repeated epochs of 100. Once the CNN is well trained, the learned network can be used to detect the salient objects of the test images.

3.3. Textural contrast CNN

There is no doubt that color cues are essential for saliency detection. However, the color features will not work well for the images shown in Figure 6 in which colors are not dominant. The salient objects in these images have uniform textural features or obvious textural contrast with the background, which reminds us to construct the textural contrast CNN. Gabor filters are defined to simulate the performance of the human visual cortical cells by decomposing an image into multiple orientations and scales. They are highly sensitive to local features, which is helpful for the extracting of discriminative textural features. In our implementation, we use the Gabor filter response with 6 orientations and 4 scales. Therefore, 24 filter responses are obtained for a given image as shown in Figure 7. These responses cannot be used as textual features because of their huge dimensions. Average filter responses within the image is computed and result in 24-dimensional textual feature.

Because the input of the Gabor feature extraction function must be regular patches instead of irregular superpixels, we feed the function with the grid that corresponds with the superpixel. The size of the Gabor global contrast cube of each grid is $24\times24\times24$. As shown in Figure 8, the textural contrast CNN is built in the same way as the color one with different convolutional layer settings. The dimensions of the four convolutional layers of the Gabor CNN are $5\times5\times96$, $3\times3\times128$, $3\times3\times64$ and $2\times2\times2$, respectively.

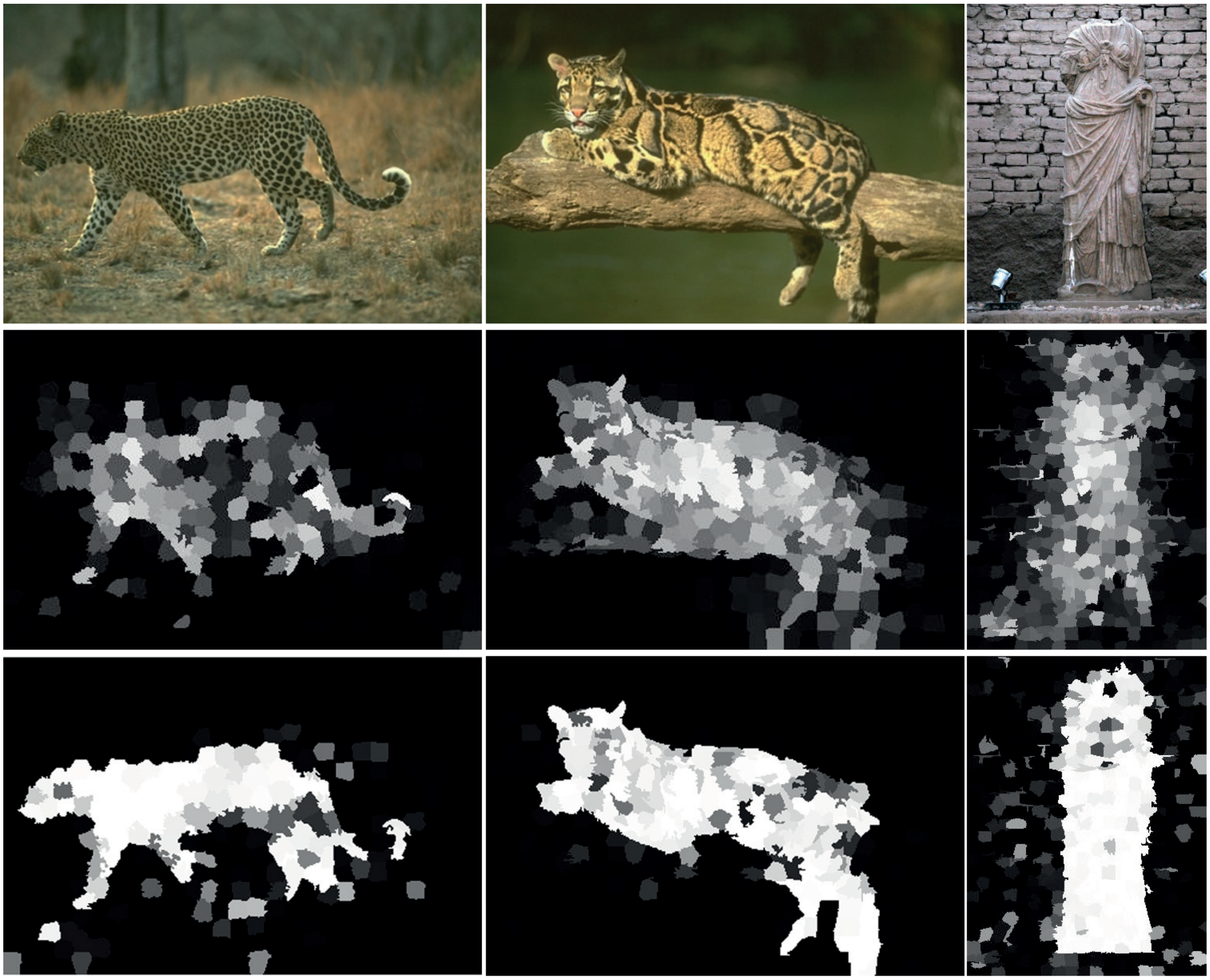


Figure 6. Images with obvious textural features. The first row shows original images, the second row shows saliency maps generated by the color contrast CNN, and the third row shows the saliency maps generated by the textural contrast CNN.

3.4. Saliency fusion

Basically, there are two fusion methods, linear weighted fusion and exponential weighted fusion, to combine saliency maps generated by different cues together. Linear weighted fusion with fixed weight applies to the situation that saliency cues are with high relativity, for example, the saliency maps of multiple scales. Exponential weighted fusion is appropriate for the diverse saliency cues, where the weights of different cues are controlled by β. Instead of using a fixed β for all the images, we use a different weight for each superpixel according to the ratio between a separated cue (color or texture) contrast and the total contrast (color and texture) within the superpixel. By this way, the higher contrast cue is highlighted in the final map while the

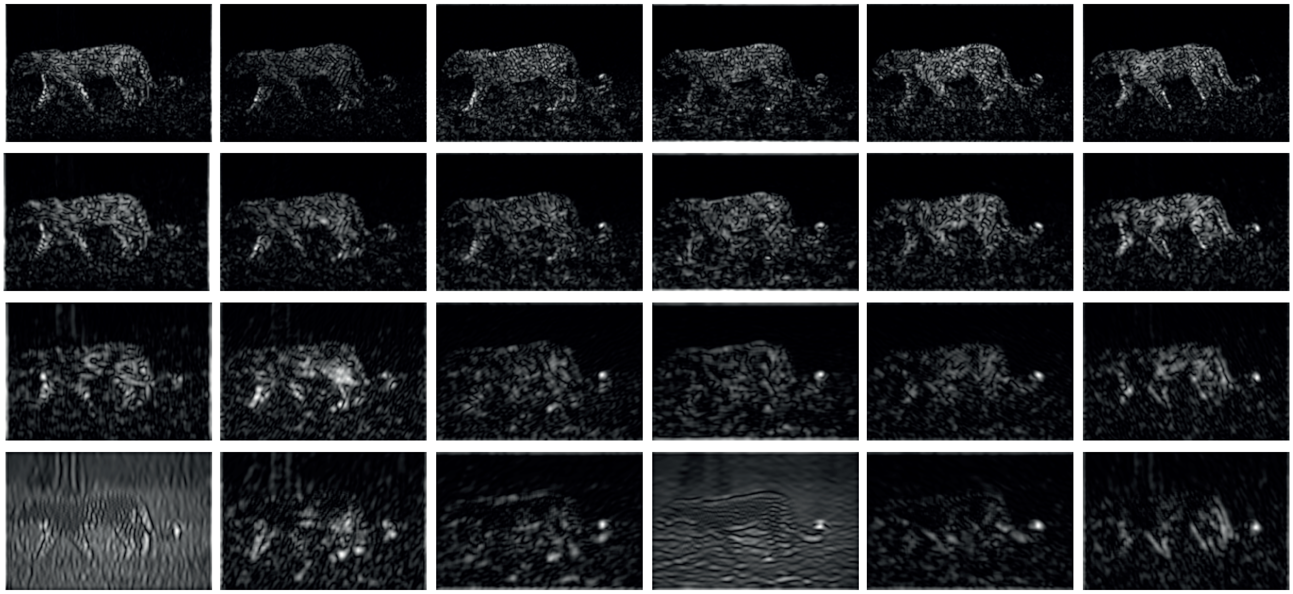


Figure 7. Gabor features of the input image.

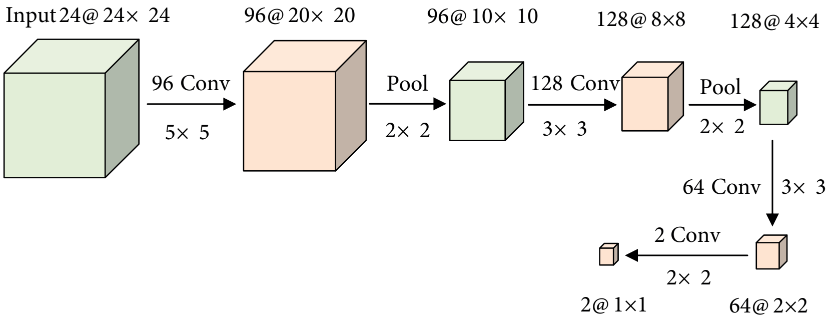


Figure 8. The structure of the textual-contrast CNN.

other one is suppressed, which facilitates achieving an optimal result. Several visual samples of saliency fusion are given in Figure 9.

$$sal(sp_i) = csal(sp_i)^{\beta} tsal(sp_i)^{1-\beta}, \tag{4}$$

where $csal$ and $tsal$ are saliency maps generated by color and textural CNN, respectively. $\beta = Cdc(sp_i)/(Cdc(sp_i)+Cdt(sp_i))$ is the proportion of color contrast of the total contrast within sp_i. For $N = 24$, color contrast $Cdc(sp_i)$ and texture contrast $Cdt(sp_i)$ of $sp_i$are given by the following:

$$Cdc(sp_i) = \sum_{j=1}^{N\times N} (cn(sp_i) - cn(sp_j)) \tag{5}$$

$$Cdt(sp_i) = \sum_{j=1}^{N\times N} (gabor(sp_i) - gabor(sp_j)) \tag{6}$$

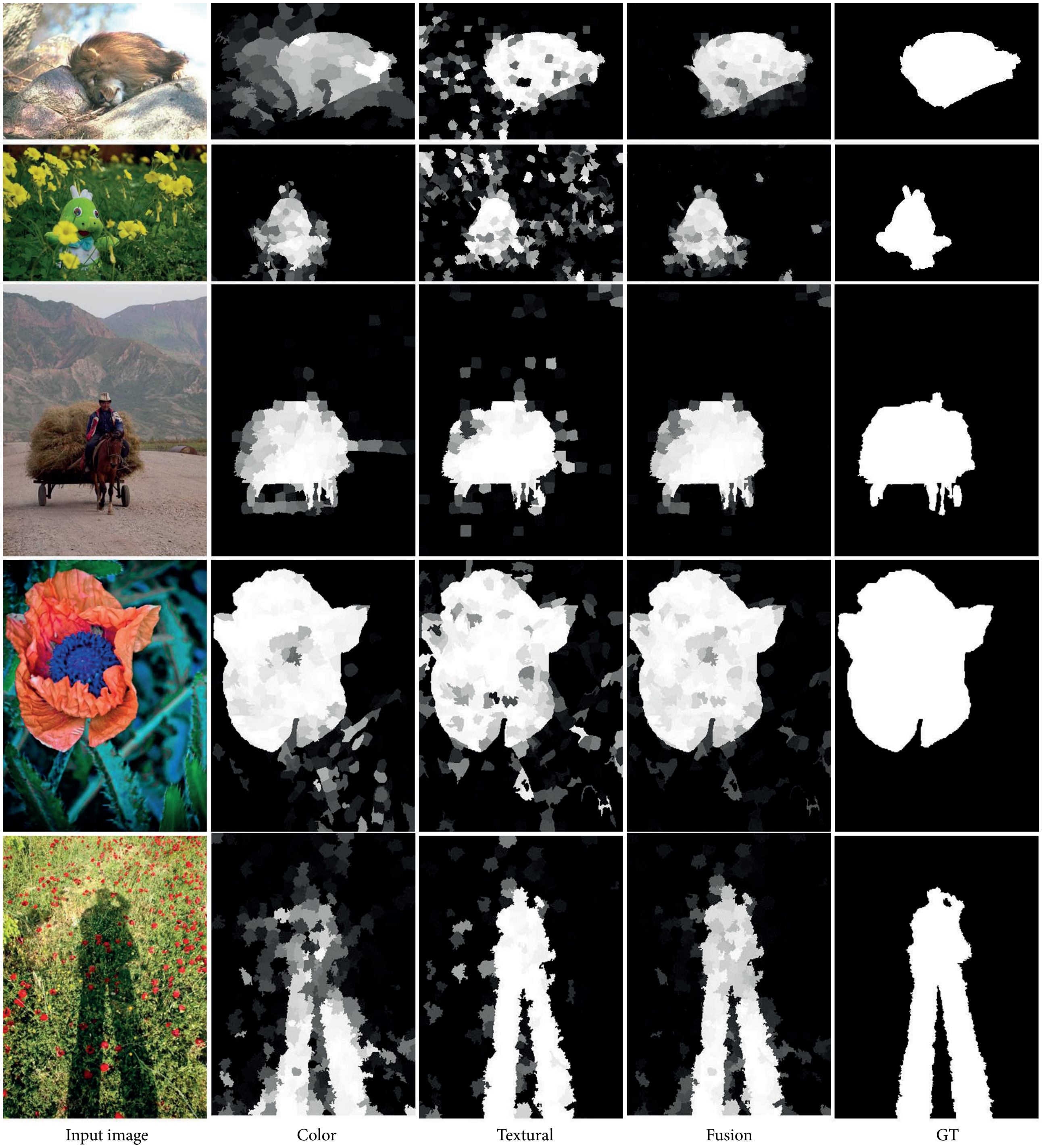


Figure 9. Visual samples of fusion saliency. The first column shows input images; the second, third, and fourth columns show saliency maps generated by color contrast CNN, textual contrast CNN and saliency fusion, respectively; the last column is the ground truth.

4. Experiment and discussion

4.1. Experimental datasets and setting

To evaluate the performance of the proposed saliency detection method, we performed a detailed experimental study on three public datasets: MSRA1000[13], ECSSD[16], and PASCAL-S[30]. The MSRA1000 dataset contains 1000 color-dominant, single-object images. The dataset is widely used as a benchmark by saliency detection methods. The ECSSD dataset is the extended version of CSSD, which covers more challenging and diverse images. PASCAL-S contains 850 images with complex backgrounds and multiple salient objects. The dataset was derived from the segmentation challenge dataset PASCAL-VOC 2012. Therefore, no prior knowledge of saliency detection was used during the construction of the dataset. The three datasets contain images covering different scenes and characteristics. Furthermore, the datasets are provided with manually labelled ground truths for training, testing, and comparison.

We train our network using 10,000 images in total, among which 9000 images come from the MSRA10K dataset and 1000 images come from the PASCAL-VOC dataset. Note that there is no overlap between our training set and test set. The diverse training samples will help the deep networks to learn more universal and robust features for natural scene images. From each image, we chose 30 salient samples and 70 nonsalient samples at random, and then approximately 1 million training samples were produced. The jitter was utilized to prevent overfitting, including horizontal reflection, rescaling, and translation.

The parameter setting during the training of our CNNs is as follows. The SGD is used for training with the initial learning rate of 0.01 and momentum of 0.9. The learning rate decreases by a ratio of 0.1 until the loss converges. The weight decay is set as 0.0005 to reduce overfitting. The batch size is set to 256 and the learning process will be iterated for 100 epochs. The more detailed description of our CNNs are listed in Tables 1 and 2. We do not list the ReLu layers here because they will not change the dimensions of outputs.

Table 1. Details of color contrast CNN. Pooling means max pooling layer; Full represents the fully connected layer; Channels means the number of output feature maps; Filter size is the size of the convolution or pooling window; Output size is the output feature map size of the layer; Neurons means the number of neurons of the layer.

Layer	Convolution	Pooling	Convolution	Pooling	Convolution	Full
Channels	64	–	96	–	48	2
Filter size	5 × 5	2 × 2	3 × 3	2 × 2	3 × 3	2 × 2
Output size	20 × 20	10 × 10	8 × 8	4 × 4	2 × 2	1 × 1
Neurons	20 × 20 × 64	–	8 × 8 × 96	–	2 × 2 × 48	1 × 1 × 2

Table 2. Details of textural contrast CNN. The items of the table are same as those in Table 1.

Layer	Convolution	Pooling	Convolution	Pooling	Convolution	Full
Channels	96	–	128	–	64	2
Filter size	5 × 5	2 × 2	3 × 3	2 × 2	3 × 3	2 × 2
Output size	20 × 20	10 × 10	8 × 8	4 × 4	2 × 2	1 × 1
Neurons	20 × 20 × 96	–	8 × 8 × 128	–	2 × 2 × 64	1 × 1 × 2

4.2. Experiment results and comparison

We compare our method with 8 state-of-the-art methods, which were proposed in the last six years and cover a variety of saliency evaluation approaches. GS [31] is a typical bottom-up method that utilizes background priors.

GMR [17] uses the graph-based representation and ranking function to evaluate the saliency of superpixels. HS[16] proposes a hierarchical framework to acquire multiscale features. wCtr[32] formulates the background weighted contrast to provide a robust measure for saliency detection. HDCT [33] evaluates a superpixel as a part of a salient object or background by mapping the RGB features into a high-dimensional feature space. MCG-GBVS [30] is a hybrid method combining object candidate generation with saliency detection. Despite different image representations and evaluation approaches, all of the five methods above are bottom-up–based saliency detection models. In contrast, the superCNN [1] transforms the issue of saliency detection into a binary classification problem that can be solved by deep learning methods. Two CNNs are formulated in LEGS [18], including one for local estimation and another for global search, to achieve superior results. We use the saliency maps provided in the authors' project sites or the results generated by the published source codes for comparison. Three commonly used measurements, including precision-recall curves, F-measures, and mean absolute errors, are utilized for the quantitative performance evaluation.

4.2.1. Precision and recall curve

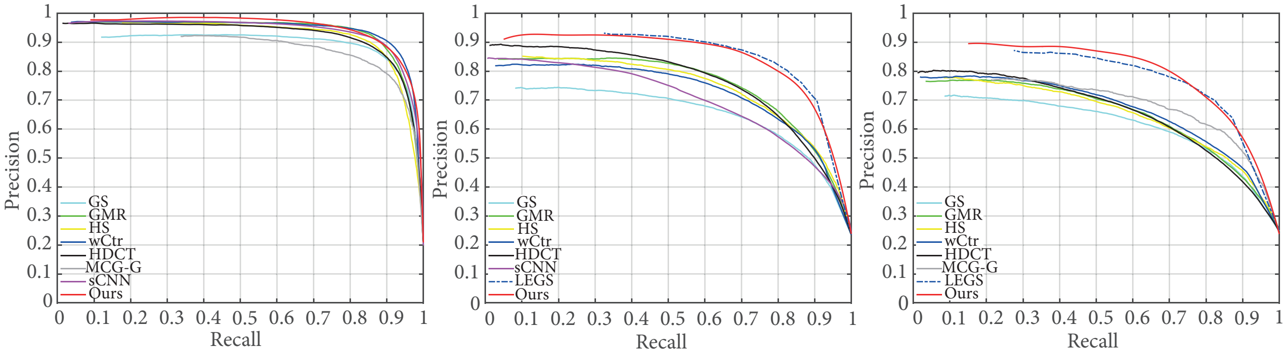
Precision is the ratio between the number of pixels that are correctly detected as salient and all the salient pixels detected by a method. It can be defined as $precision = \frac{M \cap G}{|M|}$, where M is the segmented image with the fixed threshold T, and G is the ground truth. Recall is defined as the ratio between the number of pixels that are correctly detected as salient and all the salient pixels in the ground truth mask, with the expression of $recall = \frac{M \cap G}{|G|}$. With the threshold T varying from 0 to 255, different segmented images and precision-recall pairs are generated, and a precision-recall curve is formed.

The precision-recall curves of our methods and the other seven methods on datasets MSRA1000, ECSSD, and PASCAL-S are shown in Figures 10a–10c, respectively. The minimum recall corresponds to the highest threshold. Therefore, higher precision at a lower recall value means that more salient pixels are correctly detected as parts of the object with higher confidence. Meanwhile, the higher precision at the high recall indicates that more nonsalient pixels are correctly detected as background, even with a very low threshold. When the recall value is set to 1, all the pixels in an image are considered salient. The precision at this point represents the average proportion of salient objects in the images over the entire dataset. That is why all the methods end with the same precision when using the same dataset. Although many recent state-of-the-art methods obtain satisfactory results on the widely used MSRA1000 dataset, the proposed method results in an improvement in performance. Our method achieves better performance on the complex ECSSD dataset than all the traditional salient detection methods. Compared to the other two CNN-based methods, superCNN and LEGS, our method outperforms the former and has a similar performance to the latter. Note that the result of superCNN is based on CSSD, which is a part of ECSSD. The proposed method achieves the best performance in the most challenging dataset PASCAL-S. For quantitative comparison, the proposed method achieves the maximum precision of 0.9856, 0.9250, and 0.8949 on MSRA1000, ECSSD, and PASCAL-S when the corresponding recall values are 0.3345, 0.2139, and 0.1596, which are higher than the other compared methods.

4.2.2. F-measure

F-measure is a baseline measure of detection accuracy. It is the harmonic average of the precision and recall with the following definition:

$$F = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}. \tag{7}$$



(a) (b) (c)

Figure 10. Precision-recall curves comparison on the datasets (a) MSRA1000, (b) ECSSD, and (c) PASCAL-S. Note that sCNN and MCG-G are abbreviations of superCNN and MCG-GBVS for display reasons.

An F-measure reaches its highest value at 1 and lowest at 0. Higher value means better performance of precision and recall. The salient objects usually have a small size compared to the whole image. Here we set $\beta^2 = 0.3$ to highlight the precision. For each precision-recall pair there is an F-measure value according to the definition. When the threshold T is changed from 0 to 255, an F-measure curve is generated over the whole dataset. The comparison results are given in Figures 11a–11c. Our approach is superior to the other compared methods when considering F-measure on MSRA1000 dataset. As known, MSRA1000 is a widely used single-object dataset. The bottom-up saliency methods that combine multiple saliency cues can yield good performance on the dataset. Deep learning-based methods further improve the results. The proposed method does better than all the other compared methods except LEGS on datasets ECSSD and PASCAL-S. Compared to LEGS, our method achieves higher F-measure with high threshold, which means that more foreground pixels are correctly detected as salient with higher confidence.

4.2.3. Mean absolute error

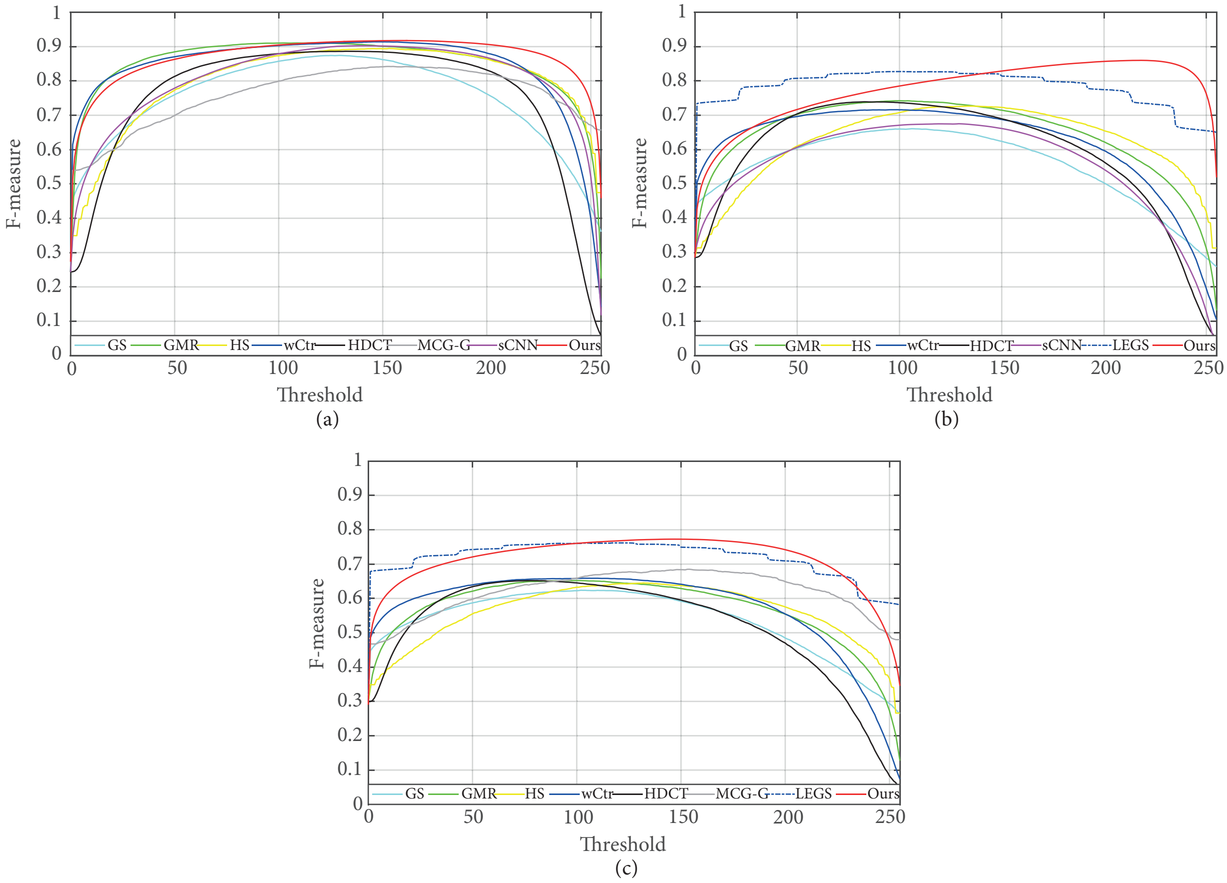
Both the precision-recall curve and F-measure focus on the pixels that are correctly detected as salient while they neglect ones that are properly detected as background. Therefore, the mean absolute error (MAE) [15] is introduced to measure the difference between the resulting saliency map and the ground truth pixel by pixel for a more faithful evaluation. The MAE is defined as follows:

$$MAE = \frac{1}{|I|} \sum_{x} |S(I_x) - G(I_x)|, \tag{8}$$

where S is a saliency map and G is the corresponding binary ground truth. x is a pixel in image I. As seen in Figures 12a–12c, our method achieves the lowest MAE value of 0.0501, 0.0929, and 0.1378 on datasets MSRA1000, ECSSD, and PASCAL-S among the compared methods. This result means that the detected saliency map of our algorithm is closest to the ground truth.

4.2.4. Component analysis

We further compare the performance of the different components of our method: the color contrast CNN and the textural contrast CNN. As shown in Figure 13, the color contrast CNN achieves better performance than the textural contrast CNN on the MSRA1000 dataset, while it is the opposite on the ECSSD dataset. This



(a)

(b)

(c)

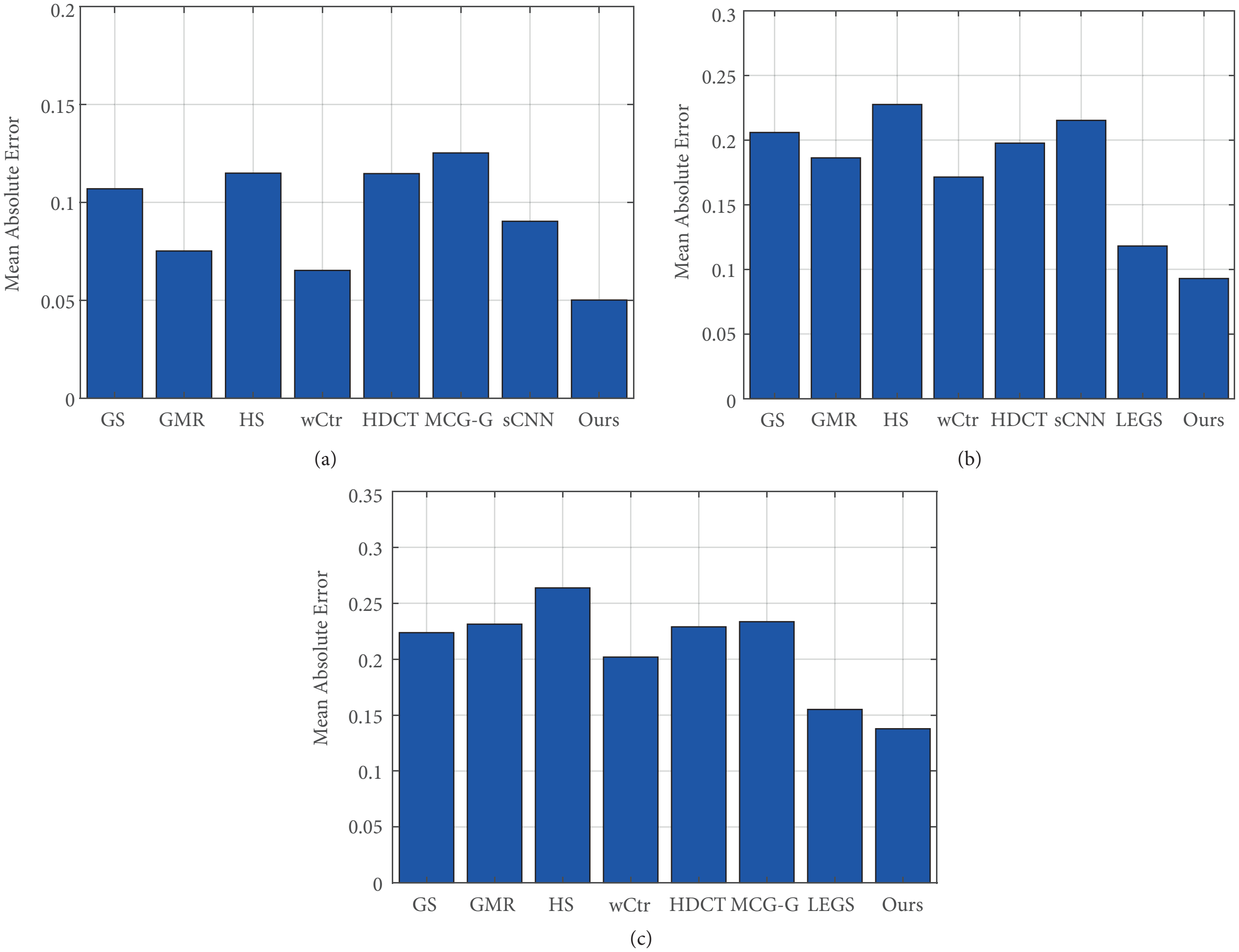
Figure 11. Fixed threshold F-measure comparison on the datasets (a) MSRA1000, (b) ECSSD, and (c) PASCAL-S. Note that the sCNN and MCG-G are abbreviations of the superCNN and MCG-GBVS for display reasons.

is because the MSRA1000 is a simple and color-dominant dataset, while more texture-dominant images are included in ECSSD.

Some visual samples of our method and other compared methods are shown in Figure 14. When human images in the first three rows are considered, some methods tend to detect the vivid clothes instead of the entire object. Our method can detect the entire human figure no matter what color of clothes they wear. The images in the fourth and fifth rows are featured by low contrast between the objects and background are a great challenge for saliency detection. Our method achieves better results than others in this case. The images in the sixth and seventh rows are featured by multiobjects, in which the less salient ones tend to be lost. The proposed method also works well for the texture-dominant image shown in the eighth row because the Gabor contrast CNN is involved. Some methods highlight the background regions instead of the object in the sample of the ninth row because of the higher color contrast of the background. For the last sample, it is difficult to detect the whole helicopter.

5. Conclusions

In this paper, we propose a deep learning-based saliency detection method in which a novel spatial-aware contrast representation is used to feed the convolutional neural networks. The spatial-aware contrast cube captures the global contrast features and preserves the spatial information of the original images. To detect



(a)

(b)

(c)

Figure 12. MAE comparison on the datasets (a) MSRA1000, (b) ECSSD, and (c) PASCAL-S. Note that sCNN and MCG-G are abbreviations of the superCNN and MCG-GBVS for display reasons.

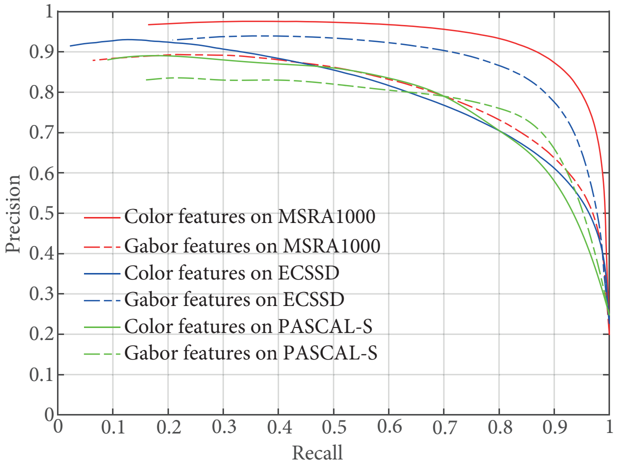


Figure 13. The comparison of the components of the proposed method.

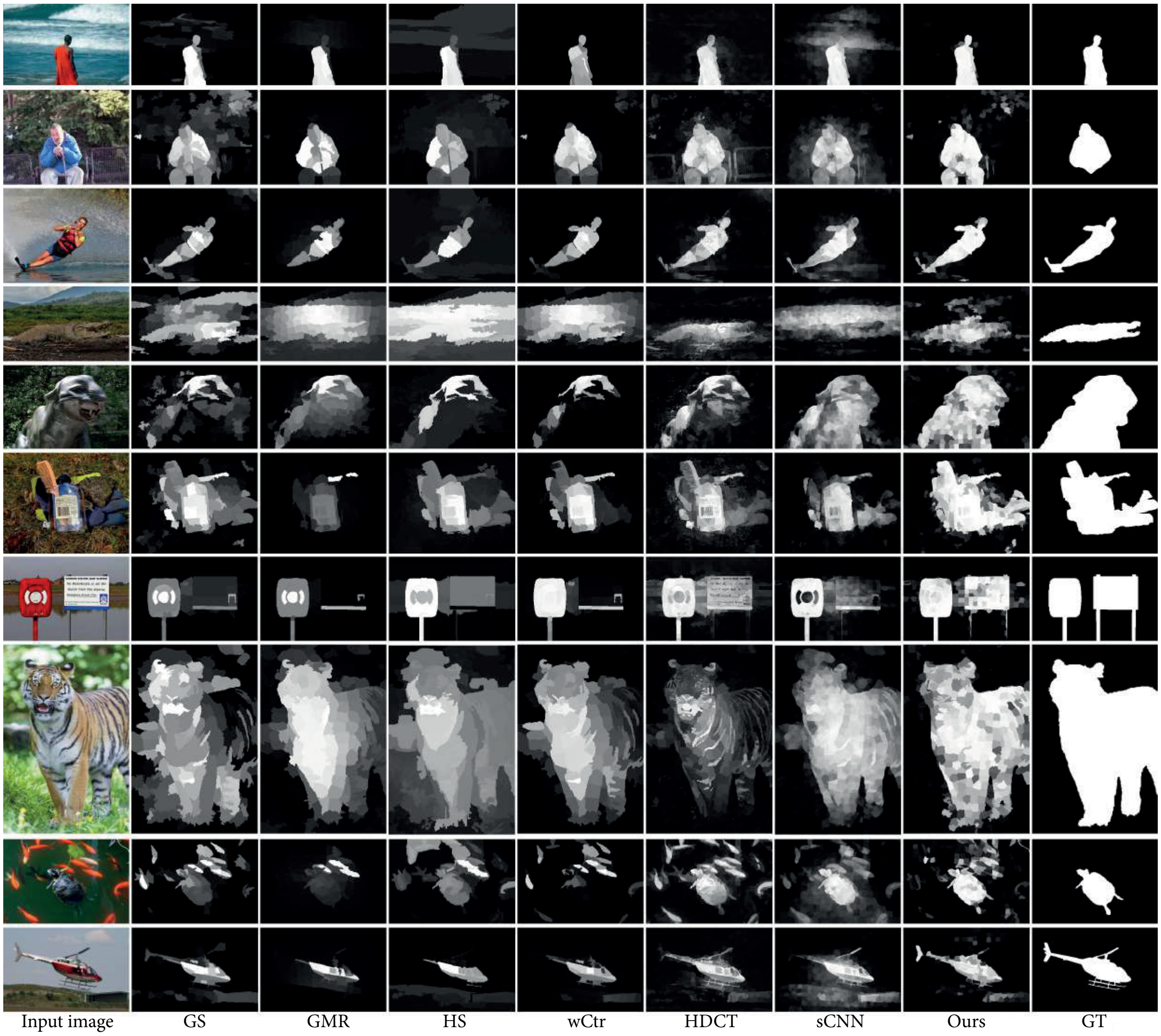


Figure 14. Visual samples of our method and other compared methods.

the salient objects from complex real scene images, we also construct the CNN, which is fed with textural features. Our method integrates spatial-aware color contrast saliency and textural contrast saliency through a dynamic fusion scheme. Experimental results on public datasets show that the proposed algorithm achieves a performance improvement, especially when the MAE measurement is concerned. Our future work will focus on the application of saliency detection in context aware image retargeting. The saliency map generated by the proposed method can be used as energy map in image retargeting, which determines the visual importance of pixels. Combining with seam carving method, the image will be retargeted with visually prominent regions preserved.

Acknowledgment

The work is supported by National Natural Science Foundation of China under Grant No.61772244, and University Natural Science Research Project of Jiangsu Province under Grant No.16KJB520009 and 15KJB520008.

References

[1] He S, Lau RWH, Liu W, Huang Z, Yang Q. Supercnn: A superpixelwise convolutional neural network for salient object detection. International Journal of Computer Vision 2015; 115(3): 330-344. doi: 10.1007/s11263-015-0822-0

[2] Cheng MM, Zhang GX, Mitra NJ, Huang X, Hu SM. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 2015; 37(3): 569-582. doi: 10.1109/TPAMI.2014.2345401

[3] Heo D, Lee E, Ko BC. Pedestrian detection at night using deep neural networks and saliency maps. Electronic Imaging 2018; 2018(17): 1-9. doi: 10.2352/J.ImagingSci.Technol.2017.61.6.060403

[4] Zhang F, Du B, Zhang L. Saliency-guided unsupervised feature learning for scene classification. IEEE Transactions on Geoscience and Remote Sensing 2015; 53(4): 2175-2184. doi: 10.1109/TGRS.2014.2357078

[5] Bai C, Chen J, Huang L, Kpalma K, Chen S. Saliency-based multi-feature modeling for semantic image retrieval. Journal of Visual Communication and Image Representation 2018; 50: 199-204. doi: 10.1016/j.jvcir.2017.11.021

[6] Shafieyan F, Karimi N, Mirmahboub B, Samavi S, Shirani S. Image retargeting using depth assisted saliency map. Signal Processing: Image Communication 2017; 50: 34-43. doi: 10.1016/j.image.2016.10.006

[7] Ramanathan S, Katti H, Sebe N, Kankanhalli M, Chua TS. An eye fixation database for saliency detection in images. In: European Conference on Computer Vision; Heraklion, Crete, Greece; 2010. pp. 30-43.

[8] Alexe B, Deselaers T, Ferrari V. Measuring the objectness of image windows. IEEE Transactions on Pattern Analysis and Machine Intelligence 2012; 34(11): 2189-2202. doi: 10.1109/TPAMI.2012.28

[9] Jiang P, Ling H, Yu J, Peng J. Salient region detection by ufo: Uniqueness, focusness and objectness. In: IEEE International Conference on Computer Vision; Sydney, Australia; 2013. pp. 1976-1983.

[10] Mangim GR. Neural mechanisms of visual selective attention. Psychophysiology 1995; 32(1):4-18. doi: 10.1111/j.1469-8986.1995.tb03400.x

[11] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998; 20(11): 1254-1259. doi: 10.1109/34.730558

[12] Liu T, Yuan Z, Sun J, Wang J, Zheng N et al. Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence 2011; 33(2): 353-367. doi: 10.1109/TPAMI.2010.70

[13] Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition; Miami, FL, USA; 2009. pp. 1597-1604.

[14] Goferman S. Context-aware saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition; San Francisco, CA, USA; 2010. pp. 2376-2383.

[15] Perazzi F, Krähenbühl P, Pritch Y, Hornung A. Saliency filters: Contrast based filtering for salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition; Providence, RI, USA; 2012. pp. 733-740.

[16] Yan Q, Xu L, Shi J, Jia J. Hierarchical saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition; Portland, Oregon, USA; 2013. pp. 1155-1162.

[17] Yang C, Zhang L, Lu H, Ruan X, Yang MH. Saliency detection via graph-based manifold ranking. In: IEEE Conference on Computer Vision and Pattern Recognition; Portland, Oregon, USA; 2013. pp. 3166-3173.

[18] Wang L, Lu H, Ruan X, Yang MH. Deep networks for saliency detection via local estimation and global search. In: IEEE Conference on Computer Vision and Pattern Recognition; Boston, MA, USA; 2015. pp. 3183-3192.

[19] Lee G, Tai YW, Kim J. Deep saliency with encoded low level distance map and high level features. In: IEEE Conference on Computer Vision and Pattern Recognition; Boston, MA, USA; 2015. pp. 660-668.

[20] Tong N, Lu H, Zhang Y, Ruan X. Salient object detection via global and local cues. Pattern Recognition 2015; 48(10): 3258-3267. doi: 10.1016/j.patcog.2014.12.005

[21] Xu L, Zeng L, Duan H. An effective vector model for global-contrast-based saliency detection. Journal of Visual Communication and Image Representation 2015; 30: 64-74. doi: 10.1016/j.jvcir.2015.03.011

[22] Xu D, Tang Z, Xu W. Salient object detection based on color names. KSII Transactions on Internet and Information Systems 2013; 7(11): 2737-2753. doi: 10.3837/tiis.2013.11.011

[23] Li G, Yu Y. Visual saliency based on multiscale deep features. In: IEEE Conference on Computer Vision and Pattern Recognition; Boston, MA, USA; 2015. pp. 5455-5463.

[24] Zhao R, Ouyang W, Li H, Wang X. Saliency detection by multi-context deep learning. In: IEEE Conference on Computer Vision and Pattern Recognition; Boston, MA, USA; 2015. pp. 1265-1274.

[25] Wang W, Shen J, Yang R, Porikli F. Saliency-aware video object segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2018; 40(1): 20-33. doi: 10.1109/TPAMI.2017.2662005

[26] Kutbay U, Hardalaç F, Akbulut M, Akaslan Ü, Serhatlıoğlu S. A Computer-Aided Diagnosis System for Measuring Carotid Artery Intima-Media Thickness (IMT) Using Quaternion Vectors. Journal of Medical Systems 2016; 40(6): 149. doi: 10.1007/s1091

[27] Bonnin-Pascual F, Ortiz A. A novel approach for defect detection on vessel structures using saliency-related features. Ocean Engineering 2018; 149: 397-408. doi: 10.1016/j.oceaneng.2017.08.024

[28] Achanta R, Shaji A, Smith K, Lucchi A, Fua P et al. SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence 2012; 34(11): 2274-2282. doi: 10.1109/TPAMI.2012.120

[29] Khan FS, Anwer RM, Van de Weijer J, Bagdanov AD, Vanrell M et al. Color attributes for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition; Providence, RI, USA; 2012. pp. 3306-3313.

[30] Li Y, Hou X, Koch C, Rehg JM, Yuille AL. The secrets of salient object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition; Columbus, OH, USA; 2014. pp. 280-287.

[31] Wei Y, Wen F, Zhu W, Sun J. Geodesic saliency using background priors. In: European Conference on Computer Vision; Firenze, Italy; 2012. pp. 29-42.

[32] Zhu W, Liang S, Wei Y, Sun J. Saliency optimization from robust background detection. In: IEEE Conference on Computer Vision and Pattern Recognition; Columbus, Ohio, USA; 2014. pp. 2814-2821.

[33] Kim J, Han D, Tai Y, Kim J. Salient region detection via high-dimensional color transform. In: IEEE Conference on Computer Vision and Pattern Recognition; Columbus, OH, USA; 2014. pp. 883-890.