Sparse Tensor Auto-Encoder for Saliency Detection

SHUYUAN YANG AND JUNXIAO WANG
School of Artificial Intelligence, Xidian University, Xi’an 710071, China

This work was supported in part by the National Natural Science Foundation of China under Grant 91438103, Grant 61771376, Grant 61771380, Grant 61703328, Grant 61438201, Grant 61701267, and Grant 61703328, in part by the Equipment Pre-research Project of the 13th Five-Years Plan under Grant 6140140050206, Grant 6140140050206, Grant 614012010103, Grant 6141A020223, Grant 6141B06160301, and Grant 6141B07090102, in part by the Major Research Plan in Shaanxi Province of China under Grant 2017ZDSXM-GY-103 and Grant 017ZDXXL-GY-03-02, in part by the National Natural Science Foundation of China, under Grant 61771376, Grant 61771376, Grant 61771380, Grant 61703328, Grant 61438201, Grant 61703328, in part by the Equipment Pre-research Project of the 13th Five-Years Plan under Grant 6140140050206, Grant 614012010103, Grant 6141A020223, Grant 6141B06160301, and Grant 6141B07090102, in part by the Foundation of the State Key Laboratory of Complex Electromagnetic Environment Effects (CEMEE) under Grant 2017K0202B, Grant 2018K0101B, Grant 2019K0203B, and Grant 2019K0201B, in part by the Science Basis Research Program in Shaanxi Province of China under Grant 16JK1823, Grant 2017JMS086, Grant 2019JQ-663, and Grant 2019JQ-663, and in part by Key Scientific and Technological Innovation Team of Shaanxi Province under Grant S2020-ZC-TD-0084.

ABSTRACT

In this paper, a new Sparse Tensor Auto-Encoder (STAE) model is proposed to learn a latent and discriminative feature representation for saliency detection. By formulating the background patches as holistic high-dimensional tensors and learning multi-dimensional dictionaries to code image patches, the coding error can precisely reveal the difference between salient object and background. Then a saliency map can be derived by a subsequent refinement of the representation errors of patches via image segmentation. Several benchmark datasets are used to verify the effectiveness of the proposed method. The results show that the proposed STAE can accurately locate saliency regions and outperform its counterparts.

INDEX TERMS

Visual saliency detection, sparse tensor auto-encoder, multi-dimensional dictionary, image segmentation, background.

I. INTRODUCTION

Visual saliency detection has received increasing interest from both computer vision and machine learning communities for decades [1]. Most of the traditional saliency models explore visual perception cues like the contrast, closure, and position priors for designing detection schemes [2]–[5]. However, these low-level cues are insufficient to model the complex pattern of visual attention. Recently, several works have been developed to establish a good representation of non-salient regions [6]–[8], which helps to distinguish the salient objects from the background. For example, Han developed stacked denoising autoencoders with deep learning architecture to model the background [6], [7]. Xia et al. [8] constructed a deep C–S inference network and trained it with the data sampled randomly from the entire image to obtain a unified reconstruction pattern for the image. Wang et al. [9] proposed a discriminative background dictionary to provide accurate saliency estimation.

In very recent years, some multi-level convolutional features are also explored to serve as the saliency cues. Li et al. [21] proposed a Multi-Task Deep Neural Network (MTDNN) model to effectively model the semantic properties of salient objects in a data-driven manner, for accurate salient object detection. In [22], the authors proposed a new salient object detection method by introducing short connections to the skip-layer structures within the Holistically-Nested Edge Detector (HED) architecture, based on a Fully Convolutional Neural Network (FCN). In [23], the authors proposed a novel bi-directional message-passing model to integrate multi-level features for salient object detection. In [25], the authors presented a framework based on multiple-instance learning, where low-, mid-, and high-level features are incorporated in the detection procedure. In [26], the authors proposed to detect salient objects based on selective contrast, in which selective contrast intrinsically explores the most distinguishable component information in color, texture, and location.

Although these works have explored the representation or coding errors for saliency detection by linear or nonlinear models [10]–[15], there are some limitations in two aspects,

1) Most of the available background representation based methods vectorize multi-view features of pixels, and then
input their features into a linear or nonlinear model for representation. However, these features are intrinsic high-order tensors, and a rough cascading and vectorization will distort the high order statistical properties of data. Thus, the representation or coding accuracy will be subsequently affected.

2) Most of the available methods work on improving the effectiveness of coding model. However, the efficiency of coding model is another important issue for saliency detection. Finding a compact or sparse coding model can locate the most significant and representative atoms, which is helpful to distinguish salient objects from background.

Tensor are the higher order generalizations of vectors and matrices, and taking full advantage of the high-order structure of data can better understand data [16], [17]. In this paper, a novel Sparse Tensor Auto-Encoder (STAE) model is proposed for saliency detection based on our previous work [9]. STAE is designed to learn the appearance model of background regions for accurate saliency measure. By formulating the background patches as holistic high-dimensional tensors and learning multi-dimensional dictionary to code patches, the coding error can precisely reveal the difference between salient object and background. Consequently, STAE can learn the intrinsic structure of background regions and less influenced by the cluttered noise in the image, and thus favors compact representation of background regions and provides a more faithful way to detect saliency. Then a saliency map can be derived by a subsequent refining of the representation errors of patches via image segmentation. The framework of the proposed saliency detection method is illustrated in Fig.1.

Compared with the available works, the contributions of our work are twofold: 1) It explores a sparse representation of feature tensors for computational saliency detection, where a latent sparse subspace is learned for background using a compact and accurate nonlinear projections. 2) An image segmentation is used to refine the representation error, to produce accurate saliency map that is more consistently distributed among perceptually uniform regions. Several benchmark datasets are used to verify the effectiveness of the proposed method. Experimental results show that the proposed STAE method can outperform its counterparts both in qualitative and quantitative evaluations.

The remainder of this paper is organized as follows. In section II, the proposed STAE is explained in detail. In section III, some experiments are taken to demonstrate the effectiveness of STAE. Section IV draws the conclusion.

II. DEEP SPARSE TENSOR AUTO-ENCODER

A. FORMULATION OF FEATURE TENSOR

In our method, we denote the image as \( I \in \mathbb{R}^{m \times n \times 3} \) and divide it into small patches \( \{i_j, i_2, \ldots, i_N\} \) with the \( k \)-th patch being \( i_k \) and \( N = (m \times n) / (p \times q) \) being the number of patches of the whole image. In our work, we formulate the feature tensor from colors, illuminations and orientations of image patches. First, the three color channels of the \( k \)-th patch are combined to form a set of color features \( i_{k}^{\text{color}} = [i_{k}^{1}, i_{k}^{2}, i_{k}^{3}] \in \mathbb{R}^{p \times q \times 3} \) with \( i_{j}^{k} \) \((j \in \{r, g, b\})\). Second, because cones are quite sensitive to illumination which can be used to explain why human attention is first captured to the regions with high illumination, then the intensity feature of the \( k \)-th patch is calculated by \( i_{k}^{\text{illumination}} = (i_{k}^{r} + i_{k}^{g} + i_{k}^{b}) / 3 \in \mathbb{R}^{p \times q \times 1} \). Third, the oriented Gabor pyramids are used to create orientation features. Here, we use the first scale of oriented Gabor pyramids in four orientations \( \theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\} \) to indicate the edge saliency of the image. The outputs of Gabor filters are denoted as \( [o_{k}^{\theta}] \), and the orientation feature of the \( k \)-th patch is calculated according to (1):

\[
\hat{i}_{k}^{\text{orientation}} = \sum_{\theta} o_{k}^{\theta} / 4 \in \mathbb{R}^{p \times q \times 4}
\]

Then we cascade the color feature, illumination feature and orientation feature along the third dimensionality to construct a feature tensor \( \hat{f}_{k} \in \mathbb{R}^{p \times q \times 8} \), where \( k = 1, \ldots, N \). The flowchart of the feature extraction is shown in Fig.2.

B. SPARSE TENSOR AUTO-ENCODER (STAE)

Image border regions are an important information source of what the background regions look like. Many sparse saliency models have used this prior to build detectors either explicitly or implicitly. However, the border regions may sometimes be corrupted by the salient foreground or image noise, which will degenerate the model performance. The work of Li et al. simply takes image border regions as a background template for sparse coding and ignores the reliability of this prior [18]. Thus, it will not work well for complex background. In this section we use the STAE to learn the intrinsic structure of data and meanwhile keep robust to outliers and noise. Thus, it is used to learn the representative information of background.
from the border regions. In this way, the learned network can be used as a detector to evaluate the saliency of each image pixel.

First we select the border pixels \( \{ i_1, i_2, \ldots, i_n \} \) to build a training set whose feature vectors are denoted as \( \mathbf{X} = \{ \mathbf{f}_k^Q \}_{k=1}^Q \). Then we construct a new sparse tensor autoencoder for coding the feature tensor \( \mathbf{f}^k \) for the \( k \)-th image patch. The encoder is a three-layer neural network with an encoding layer and decoding layer. In the encoding layer, we map the feature tensor \( \mathbf{f}^k \) to a hidden representation tensor \( \mathbf{y}^k \) using an affine transformation with a set of dictionaries or projections \( \{ \mathbf{W}_1^k, \mathbf{W}_2^k, \mathbf{W}_3^k \}_{k=1}^Q \) and a bias tensor \( \mathbf{b} \), where \( \mathbf{W}_1^k \in \mathbb{R}^{p \times p}, \mathbf{W}_2^k \in \mathbb{R}^{q \times q}, \mathbf{W}_3^k \in \mathbb{R}^{1 \times 8} \), \( \mathbf{b}^k \in \mathbb{R}^{p \times q \times 8} \) and \( L \) is the number of neurons. The Sigmoid function is used as the activation function of the coding layer, and the output of the \( l \)-th neuron is,

\[
\mathbf{y}^k_l = \text{sigm}(\mathbf{f}^k \times_1 \mathbf{W}_1^k \times_2 \mathbf{W}_2^k \times_3 \mathbf{W}_3^k + \mathbf{b}^k) \quad (2)
\]

to obtain the output tensor of the hidden layer \( \mathbf{y}^k \in \mathbb{R}^{p \times q \times 1} \). The subscript \( \times_q(q = 1, 2, 3) \) is an \( n \)-mode product of the tensor with a dictionary \( \mathbf{W}_q^l \) [16], [17]. Then a decoding layer is adopted to map \( \mathbf{y}^k \) to a representation tensor \( \mathbf{z}^k \) through a sigmoid transformation,

\[
\mathbf{z}^k = \text{sigm} \left( \sum_{l=1}^L \mathbf{y}^k_l \times_1 \mathbf{W}_1^l \times_2 \mathbf{W}_2^l \times_3 \mathbf{W}_3^l + \mathbf{b}^k \right) \quad (3)
\]

where \( \mathbf{b}^k \in \mathbb{R}^{p \times q \times 8} \) is the bias tensor in the decoding layer. In the coding layer and decoding layer, a same set of neurons in the hidden layer is smaller than that of the input and output layer, so we call the autoencoder as sparse autoencoder. The structure of the proposed STAE is shown in Fig.3.

By setting the desired output of the network as \( \mathbf{t}^k \), we can train the STAE via the optimization of a set of sparse coding parameters \( \theta = \{ \mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{b}, \mathbf{b}^k \} \). The mean-squared reconstruction error between the input tensor \( \mathbf{X} = \{ \mathbf{f}_k^Q \}_{k=1}^Q \) and output tensor \( \mathbf{Z} = \{ \mathbf{z}_k^Q \}_{k=1}^Q \) is minimized via

\[
\arg \min_{\theta} L(\mathbf{X}, \mathbf{Z}) = \arg \min_{\theta} \sum_{k=1}^Q \left\| \mathbf{t}^k - \mathbf{z}^k \right\|_2^2 / 2 \quad (4)
\]

where \( \left\| \cdot \right\|_2^2 \) is the squared sum of each element in the tensor. In our method, we use the gradient descent algorithm to update the weights and bias [17]. Consequently the difference between the training data and their reconstructed data is minimized. In Fig.4, we display some learned dictionaries \( \mathbf{W}_1^k \) and \( \mathbf{W}_2^k \). From it we can observe that the learned dictionaries appear to be structured and informative and different neurons in the hidden layer will response to certain types of features. The connected weights of the hidden neuron in the third row and fifth column have strong response to the color and will only be activated when the corresponding pattern appear in the input signal. Also, the weights of the hidden neuron in the fourth row and third column are sensitive to orientation information and thus orientation information in the background regions can be captured in this way. One merit of this sparse tensor autoencoder network is that it can learn intrinsic features of image background regions without being influenced by outliers and noises. Thus, the trained network will possess the discriminative capability to fulfill the saliency detection task. Besides, it is also worthwhile to mention that the learning process will converge quickly after a few iterations.

**C. SALIENCY MAP GENERATION FROM RECONSTRUCTION ERRORS**

From a machine learning perspective, the above process tries to learn a nonlinear mapping function from noisy data that can produce sparse latent representations. The learned model can fit well with the background regions if it is properly trained. Since foreground regions does not follow the same distribution with background regions, they will be grossly irregular signals and not match with the trained model. Therefore, it is reasonable to connect the reconstruction error of each region with its probability of saliency. For regions from the background, their reconstruction errors will be small and so are their saliency values. On the contrary, regions from the foreground will share large saliency values due to the enormous reconstruction errors presented.

Given that model parameters of the sparse tensor autoencoder after training are \( \hat{\theta} = \{ \hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2, \hat{\mathbf{W}}_3, \hat{\mathbf{b}}, \hat{\mathbf{b}}^k \} \), the saliency of each patch is defined as

\[
S_i = \left\| \mathbf{f}^k - \mathbf{z}^k \right\|_2^2 \quad (5)
\]
where $s_j$ is the saliency value of the $i$-th image patch, which is the reconstruction error of its corresponding feature vector under the trained sparse tensor autoencoder network. The STAE network can be seen as a tool for dimensionality reduction. The original feature space is non-linearly mapped to a new space where the true background regions lie in. Thus we can use the reconstruction error of each patch under the learned background dictionary to measure its saliency. In our work we further process the error map to obtain visually appealing high quality saliency map. The mean-shift segmentation algorithm [19] is used to divide the original image into several uniform regions $\{r_1, r_2, \ldots, r_k\}$ and then use the following rule to adjust the error map,

$$s_{r_i} = \frac{1}{|S_{r_i}|} \sum_{j=1}^{n} n(r_i, p_j) \times s_{p_j} \quad (6)$$

where $p_j$ and $s_{p_j}$ are the $j$-th patch and its reconstruction error, $n(r_i, p_j)$ is the number of overlapping pixels between $i$-th superpixel and $j$-th patch, and $|S_{r_i}|$ represents the cardinality of the $i$-th superpixel $r_i$. After this step, the saliency values are more consistently distributed among perceptually uniform regions. Finally, we use a low pass Gaussian filter to smooth the above result to deliver better visual effect. We choose a $3 \times 3$ window with the variance of the Gaussian filter being set to 0.5. In this way, the quality of the generated saliency maps can be improved.

### III. EXPERIMENTAL RESULTS

#### A. EXPERIMENTAL SETUP

In this section several benchmark datasets are used to investigate the performance of STAE. The parameters of the proposed model include: 1) the patch size $p \times q$, 2) the number of hidden neurons in the network, $L$, and 3) the maximum number of iterations in the training stage. By extensive experiments, we carefully choose the above parameters according to the empirical analysis. For the patch size, we set $p$ and $q$ both to be 8 to make a compromise between model efficiency and effectiveness. As to the number of hidden neurons $L$, it is shown that 25 will be a good choice by the experimental trial, with which we can learn good features for discriminative saliency detection. The maximum number of iterations is set to be 100 to ensure the convergence of the objective function with small time cost.

Some related works are chosen for comparison: 1) Context-aware (CW)-[2]; 2) Dense and sparse reconstruction (DSR)-[18]; 3) Frequency-tuned (FT)-[4]; 4) Graph-based Visual Saliency (GBVS)-[5]; 5) Low-Rank Matrix Recovery (LRMR)-[24]; 6) Deep Saliency (DS)-[21]; 7) Deeply supervised salient object detection (DSS)-[22] and 8) Bi-Directional Message Passing Model (BMPM)-[23] based saliency detection methods.

Two well-known datasets, MSRA-1000 and ECSSD are used to investigate the performance of the proposed STAE. To make quantitative evaluation of different methods, several numerical indexes are employed for objective assessment. For all the comparison methods, we use the codes provided by the authors. The parameters in the original codes are also adopted in the comparative methods.

#### B. EXPERIMENTAL RESULTS ON THE MSRA-1000 DATASET

In this section we first test our method on the MSRA-1000 dataset. The detection results of some images in the dataset by the nine methods are shown in Fig.5. In a top down manner, each row in the figure corresponds to the original image, saliency maps by CW [2], DSR [18], FT [4], GBVS [5], LRMR [24], DS [21], DSS [22], BMPM [23], our proposed STAE and the ground-truth.

As we can see from the results, the saliency maps by STAE are generally better than most of the comparative methods, including DSR and several deep learning based methods that rely on the vector or matrix operations. In the comparative methods, DSR is quite related to our method in several aspects and it achieves satisfactory results on this dataset. The background regions in the saliency maps of DSR are well suppressed and it can also stress the complete object region. However, simply depending on the border templates for dense and sparse reconstruction will somewhat fail to capture the intrinsic structure of background regions. Since our proposed STAE method exploits the sparse autoencoder model for background information mining, it can get more representative and discriminative saliency cues for detection. Consequently we can see from the figure that...
the saliency maps produced by STAE can indicate interest regions accurately and meanwhile have high foreground-background contrast.

Similar to our method, LRMR also adopts the sparse idea for saliency modeling. However, the background regions are not well suppressed in the saliency maps of LRMR and also the foreground regions are not uniformly highlighted. This is due largely to the strict assumptions cast by the author on the problem. In many situations, it is difficult to find a latent space under which the foreground regions are sparse and background regions are low-rank. From this point, our model is more likely to work well because of the non-linear mechanisms. For other comparative methods, CW tends to highlight the object edges and fails to detect the entire object region. In salient object detection, it is desired that the full object region be segmented for semantic scene understanding. Therefore, our method will be more suitable for this purpose. FT tends to highlight regions whose colors deviate from the norm. As is believed, saliency is generated in a complex way and simply depending on this mean subtraction operation is not sufficient enough. This can be revealed in the saliency maps of FT where the large salient regions are sometimes assigned with small saliency values. GBVS works in a down-sampled scale mainly for computational efficiency. However, the boundaries of the salient object regions are clearly distinguished and lower the visual observation quality. Compared with the deep learning methods including DS, DSS and BMPM methods, SATE can explore the high order structure of data and avoid the rough cascading and vectorization, thus achieving more efficient and plausible representation or coding. Consequently, the comparative deep learning based methods prone to loss some details in the coding of object or background regions. Therefore, compared with other methods, STAE can produce more visually appealing saliency maps.

C. EXPERIMENTAL RESULTS ON THE ECSSD DATASET

In this section the ECSSD dataset [20] is further used to test the performance of our proposed STAE. The detection results of some images in the dataset by the nine methods are shown in Fig.6. In a top down manner, each row in the figure corresponds to the original image, saliency maps by CW [2], DSR [18], FT [4], GBVS [5], LRMR [24], DS [21], DSS [22], BMPM [23], STAE and the ground-truth. As the results show, our method can present the comparable results with the deep learning based methods. Compared with other linear representation models, our proposed method performs the best under cluttered, low contrast and multiple salient object cases, which further confirms the effectiveness of the proposed sparse tensor autoencoder based saliency detection method in real world problems. For example, we can also observe from Fig.6 that DSR tends to highlight only one salient object while ignores the others in multiple object situations (see the third and sixth images as an example). This indicates that object bias based refinement is not suitable for many real world problems. Conversely, our work is inspired by the deep and sparse cognition characteristics of human visual processing mechanism and can model the complex pattern of visual attention, thus achieving more accurate detection results when compared with the linear representation models.

D. NUMERICAL RESULTS ON THE TWO DATASETS

To make quantitative evaluation of different methods, several numerical indexes are employed to assess their performance objectively. We record the precision, recall and F-measure indexes of all the saliency methods on the datasets. Fig.7(a) shows the average PR curves of the nine saliency models on the MSRA-1000 dataset. From it we can see that the curve of STAE keeps lying to the top-right most of the plot, indicating better performance than the other methods. This is also in accordance with the above qualitative evaluation results in section III.C. Meanwhile, it can be seen from the figure that DS [21], DSS [22], BMPM [23], DSR [18] and LRMR [24] are very competitive saliency models, which further verifies the close connection between sparsity and saliency. Our model is also based on sparse characteristics of visual saliency but it is deep and nonlinear. This explains the outstanding performance of the proposed method. Fig.7(b) shows the average PR curves of the nine saliency models on the ECSSD dataset. From it we can see that the three deep learning models, STAE, DS and BMPM present comparable results for this dataset. For introducing skip
connections and deep layers, DDS present the best results. The curve of DS lies in the top-right most of the plot in Fig.7(b).

Moreover, we also compare the average time consumption of various saliency methods on MSRA-1000 dataset and ECSSD dataset. Thirty experiments are taken independently and the running time of different methods are calculated. The average consumed time of nine methods are shown in the Table 1. From the figure we can observe that the consumed time for the two datasets are similar because the images in the two datasets have the same size and thus with nearly equal computational cost. From the comparative methods we can see that our proposed STAE method has rapider implementation when compared with DSS [22], DS [21], BMPM [23], LRMR [24] and CW [2]. Thus, our method has lower computational cost when compared with most of the comparative methods.

**TABLE 1.** The consumed time of different methods.

| DATA   | CW | DSR | FT | GBVS | LRMR | STAE | DS | DSS | BMPM |
|--------|----|-----|----|------|------|------|----|-----|------|
| MSRA   | 73.42 | 6.26 | 0.19 | 0.47 | 18.85 | 1.36 | 127.70 | 99.24 | 110.35 |
| ECSSD  | 76.60 | 0.37 | 0.16 | 0.71 | 36.39 | 1.64 | 149.38 | 108.18 | 110.66 |

Furthermore, we segment each saliency map with a threshold, which is two times the mean saliency value of the map, to get a binary map. This binary map is then compared with the ground-truth map to obtain the precision, recall and F-measure values. All the precision, recall and F-measure values of six methods on the MSRA-1000 dataset are averaged and plotted in Fig.8(a).

As shown in Fig.8(a), STAE achieves the highest precision, recall and F-measure scores among all the comparison methods. It is generally accepted that a good saliency model should keep both high precision and recall indexes. Some models like FT and GBVS either obtain high precision or recall but at the expense of suffering from the other one. Different from them, our model can make a good compromise between the above two complementary indexes, thus more competitive for saliency detection. Similarly, we show the PR curves and mean precision, recall and F measure indexes of different methods on the ECSSD dataset in Fig.8(b) for quantitative evaluation. From it we can see that STAE also ranks the top 4 in mean precision, recall and F measure.

**IV. CONCLUSION**

In this paper a new Sparse Tensor Auto-Encoder (STAE) model is proposed to learn latent feature representations for discriminative saliency detection. To verify the effectiveness of the STAE, we test it on two benchmark datasets, and compare it with some related works. Both quantitative and qualitative evaluations are conducted and experimental results show that our method can produce visually high quality.
saliency maps. In the future, we will extend this model to include a deeper architecture and abundant structure (such as attention, skip connection, pyramid and so on) for more reliable saliency detection.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[2] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[3] N. Tong, H. Lu, L. Zhang, and X. Ruan, “Saliency detection with multi-scale superpixels,” IEEE Signal Process. Lett., vol. 21, no. 9, pp. 1035–1039, Sep. 2014.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in Proc. CVPR, Jun. 2009, pp. 1597–1604.

[5] B. Schölkopf, J. Platt, and T. Hofmann, “Graph-based visual saliency,” in Proc. NIPS, 2007, pp. 545–552.

[6] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, “Background prior-based salient object detection via deep reconstruction residual,” IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 8, pp. 1309–1321, Aug. 2015.

[7] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, “Advanced deep-learning techniques for salient and category-specific object detection: A survey,” IEEE Signal Process. Mag., vol. 35, no. 1, pp. 84–100, Jan. 2018.

[8] C. Xia, F. Qi, and G. Shi, “Bottom-up visual saliency estimation with deep autoencoder-based sparse reconstruction,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 6, pp. 1227–1240, Jun. 2016.

[9] S. Wang, M. Wang, S. Yang, and K. Zhang, “Salient region detection via discriminative dictionary learning and joint Bayesian inference,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 5, pp. 1116–1129, May 2018.

[10] E. Rahlu, J. Kannala, M. Salo, and J. Heikkilä, “Segmenting salient objects from images and videos,” in Proc. ECCV, 2010, pp. 366–379.

[11] J. Zhang and S. Sclaroff, “Exploiting surroundedness for saliency detection: A Boolean map approach,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 5, pp. 889–902, May 2016.

[12] L. Huafeng, L. Jing, Z. Peiyun, L. Dachuan, and L. Dongmin, “Saliency detection using adaptive background template,” IET Comput. Vis., vol. 11, no. 6, pp. 389–397, 2017.