DEEP NETWORKS BASED ENERGY MODELS FOR OBJECT RECOGNITION FROM MULTIMODALITY IMAGES

Faculty of Engineering & Information Technologies

The University of Sydney

A thesis submitted in fulfillment of the requirements for the degree of Master of Philosophy in the School of Information Technologies at The University of Sydney

Ke Yan

September 2016

Abstract

Object recognition has been extensively investigated in computer vision area, since it is a fundamental and essential technique in many important applications, such as robotics, auto-driving, automated manufacturing, and security surveillance. According to the selection criteria, object recognition mechanisms can be broadly categorized into object proposal and classification, eye fixation prediction and saliency object detection.

Object proposal tends to capture all potential objects from natural images, and then classify them into predefined groups for image description and interpretation. For a given natural image, human perception is normally attracted to the most visually important regions/objects. Therefore, eye fixation prediction attempts to localize some interesting points or small regions according to human visual system (HVS). Based on these interesting points and small regions, saliency object detection algorithms propagate the important extracted information to achieve a refined segmentation of the whole salient objects.

In addition to natural images, object recognition also plays a critical role in clinical practice. The informative insights of anatomy and function of human body obtained from multimodality biomedical images such as magnetic resonance imaging (MRI), transrectal ultrasound (TRUS), computed tomography (CT) and positron emission tomography (PET) facilitate the precision medicine. Automated object recognition from biomedical images empowers the non-invasive diagnosis and treatments via automated tissue segmentation, tumor detection and cancer staging.

The conventional recognition methods normally utilize handcrafted features (such as oriented gradients, curvature, Haar features, Haralick texture features, Laws energy features, etc.) depending on the image modalities and object characteristics. It is challenging to have a general model for object recognition. Superior to handcrafted features, deep neural networks (DNN) can extract self-adaptive features corresponding with specific task, hence can be employed for general object recognition models. These DNN-features are adjusted semantically and cognitively by over tens of millions parameters corresponding to the mechanism of human brain, therefore leads to more accurate and robust results. Motivated by it, in this thesis, we proposed DNN-based energy models to recognize object on multimodality images. For the aim of object recognition, the major contributions of this thesis can be summarized below:

1. We firstly proposed a new comprehensive autoencoder model to recognize the position and shape of prostate from magnetic resonance images. Different from the most autoencoder-based methods, we focused on positive samples to train the model in which the extracted features all come from prostate. After that, an image energy minimization scheme was applied to further improve the recognition accuracy. The proposed model was compared with three classic classifiers (i.e. support vector machine with radial basis function kernel, random forest, and naive Bayes), and demonstrated significant superiority for prostate recognition on magnetic resonance images. We further extended the proposed autoencoder model for saliency object detection on natural images, and the experimental validation proved the accurate and robust saliency object detection results of our model.

2. A general multi-contexts combined deep neural networks (MCDN) model was then proposed for object recognition from natural images and biomedical images.

Under one uniform framework, our model was performed in multi-scale manner. Our model was applied for saliency object detection from natural images as well as prostate recognition from magnetic resonance images. Our experimental validation demonstrated that the proposed model was competitive to current state-of-the-art methods.

3. We designed a novel saliency image energy to finely segment salient objects on basis of our MCDN model. The region priors were taken into account in the energy function to avoid trivial errors. Our method outperformed state-of-the-art algorithms on five benchmarking datasets. In the experiments, we also demonstrated that our proposed saliency image energy can boost the results of other conventional saliency detection methods.

Publications

1. **K. Yan**, C. Li, X. Wang, A. Li, Y. Yuan, J. Kim, D. Feng, "Adaptive Background Search and Foreground Estimation for Saliency Detection via Comprehensive Autoencoder", *The International Conference on Image Processing (ICIP)*, 2016, accepted and in publishing.

2. **K. Yan**, C. Li, X. Wang, A. Li, Y. Yuan, D. Feng, M. Khadra, J. Kim, "Automatic Prostate Segmentation on MR Images with Deep Network and Graph Model", *The 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, accepted and in publishing.

3. **K. Yan**, C. Li, X. Wang, Y. Yuan, A. Li, J. Kim, B. Li, D. Feng, "Comprehensive Autoencoder for Prostate Recognition on MR images", *The International Symposium on Biomedical Imaging (ISBI)*, 2016, published.

Attribution statements

Chapter 3 of this thesis is published as [1, 2]. I designed the study, conducted the experiment, analysed the data and wrote the manuscripts.

Acknowledgements

First and foremost I would like to express my gratitude to my primary supervisor, Prof. Dagan Feng, for his generous support and all-round guidance throughout my research. His broad knowledge and keen perspective to the future research trends have deeply impressed me and always inspire me for significant personal achievement.

I would like to offer my great appreciation to my associate supervisor, Dr. Xiuying Wang, for her continuous support, patient guidance and professional advice to my research work. I am also grateful for her valuable ideas and constructive inputs to my papers, which has greatly enhanced the papers both in contributions and in contents.

I would like to thank Dr. Changyang Li for his support both in research and life during my MPhil study. His research dedication and enthusiasm always motivate me for greater promotion. I am glad to work with Dr. Li.

I would also like to express my gratitude to A/Prof. Jinman Kim who provided research assistant position for me, so that I could have cooperation with Nepean Hospital for data collection and medical technique support. It is my pleasure to work with A/Prof. Kim. Besides, many thanks to A/Prof. Kim for his valuable reviews of my papers.

Furthermore, many thanks and best wishes to all staff and fellow students, especially to Ang Li and Yuchen Yuan, for their necessary supports to my research.

Last but not least, I would like to thank my parents for their supports and encouragements during my MPhil study.

Lists of figures

Lists of tables

Contents

1. Introduction

1.1. Object recognition on multimodality images

Object recognition (or object proposal), which tends to discover a set of regions containing object instances on an image, is an important task in computer vision. As a pre-process of image classification, object recognition directly affects the accuracy in many computer-aided detection system, such as face detection, pedestrian detection [3] and action recognition [4].

By narrowing the criteria of selection for objects, object recognition on natural image can be transformed as a target-driven task, such as saliency detection [5]. As one of the popular topics in object recognition, saliency detection is to softly recognize the most informative regions or objects corresponding to the human visual system (HVS), thus can facilitate a wide range of multimedia applications (e.g. image resizing [6] and image montage [7]). Saliency detection is composed of two sub-areas: saliency fixation prediction and saliency object detection. While saliency fixation prediction focuses on human fixation locations, saliency object detection tends to recognize the whole meaningful regions. In this thesis, we validate our proposed models on natural image to investigate saliency object detection.

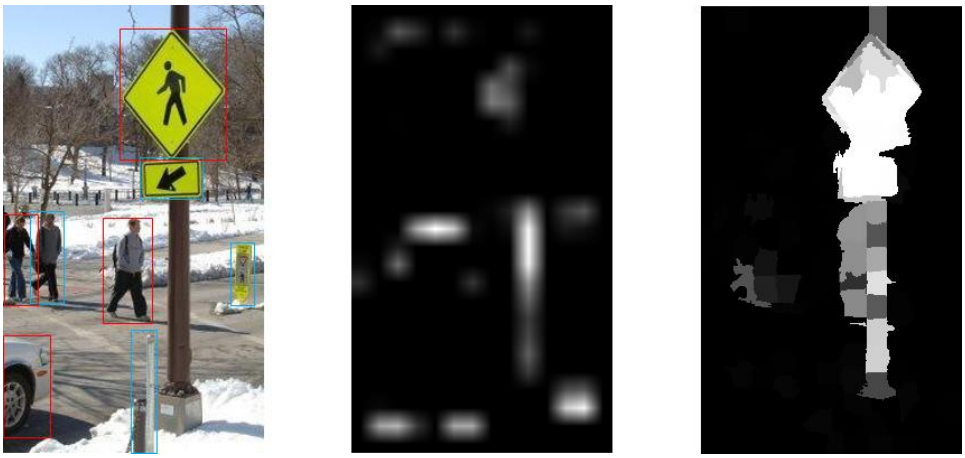


Figure 1.1 From left to right: object proposal, saliency fixation prediction by [5] and saliency object detection by [8].

Object recognition algorithm can also be applied on biomedical imaging for computer-aided diagnosis (CAD). With the visual depiction of the interior of human body, medical imaging is a necessary and effective tool for disease diagnosis and treatment. Various imaging modalities have been widely applied in clinical practice, such as magnetic resonance (MR) imaging, transrectal ultrasound (TRUS) and computed tomography (CT). As the serious diseases (e.g. prostate cancer [9], heart disease [10] and Alzheimer's disease [11]) greatly threaten the public health, it desirably requires reliable computer-aided analysis for medical imaging to improve the efficiency of clinical treatment.

In clinical practice, the computer-aided analyses include object recognition, image segmentation, tumor classification, cancer staging, etc. As an essential prerequisite for other imaging analyses, object recognition is a fundamental step in disease diagnosis. To approximate the position of objects, object recognition on medical image is usually tissue-driven, corresponding to the specific anatomical structure to be further treated. In this thesis, we also validate our proposed models to recognize prostate on MR images.

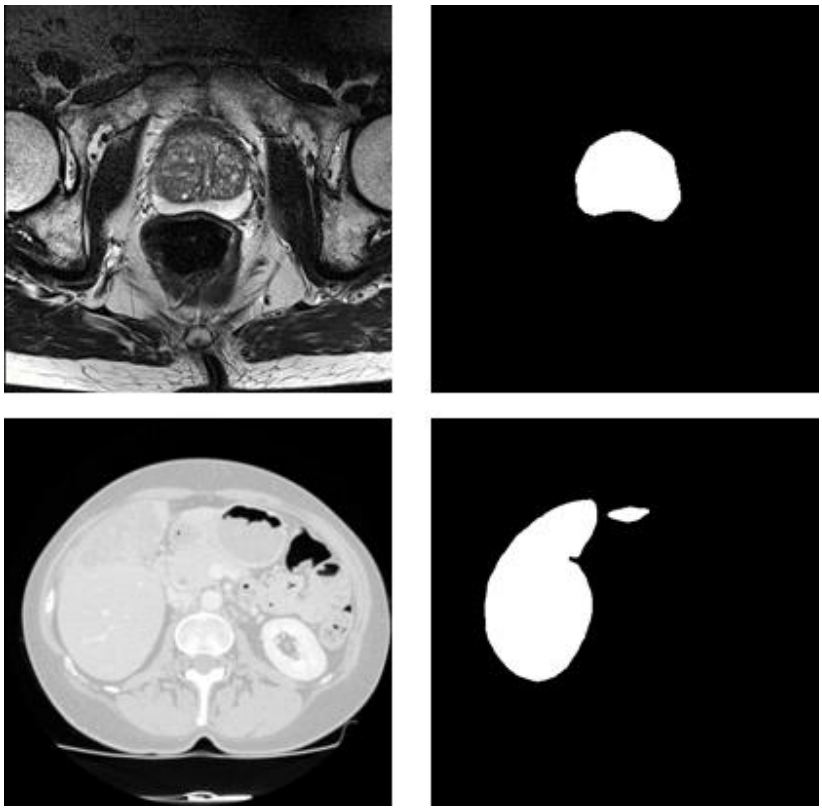


Figure 1.2 From top to bottom: MR prostate image and corresponding human annotation, CT liver image and corresponding human annotation.

1.2. Challenges

Object recognition poses some serious challenges, both on natural images and medical images. For saliency object detection on natural images, inappropriate priors and complex image scenarios usually impede precise recognition of salient objects. The conventional methods cannot extract the whole salient objects from complex image backgrounds and even generate ‘inverse’ results when using inappropriate priors. On medical images, the various artifacts make it more difficult to accurately locate the aimed anatomical structures.

1.2.1. Inappropriate priors

A common approach for saliency object detection is to select several background seeds as the first step and then to apply various strategies to form the saliency map, such as cellular automata [12], manifold ranking [13, 14], bootstrap learning [15], Markov chain [16, 17], normalized cut [18], and foreground connectivity [19]. The background seeds selection thus is an essential step and directly affects the accuracy of the saliency detection. However, most existing methods [12, 15, 17, 20] simply use image boundaries as the background seeds. Such boundary-background seed selections are technically sound for simple image sets (e.g. MSRA-10K [21]), but are at risk of failing to produce saliency map for complex image sets (e.g. ECSSD [22] and PASCAL-S [23]) when the candidate objects are attached to the image boundaries.

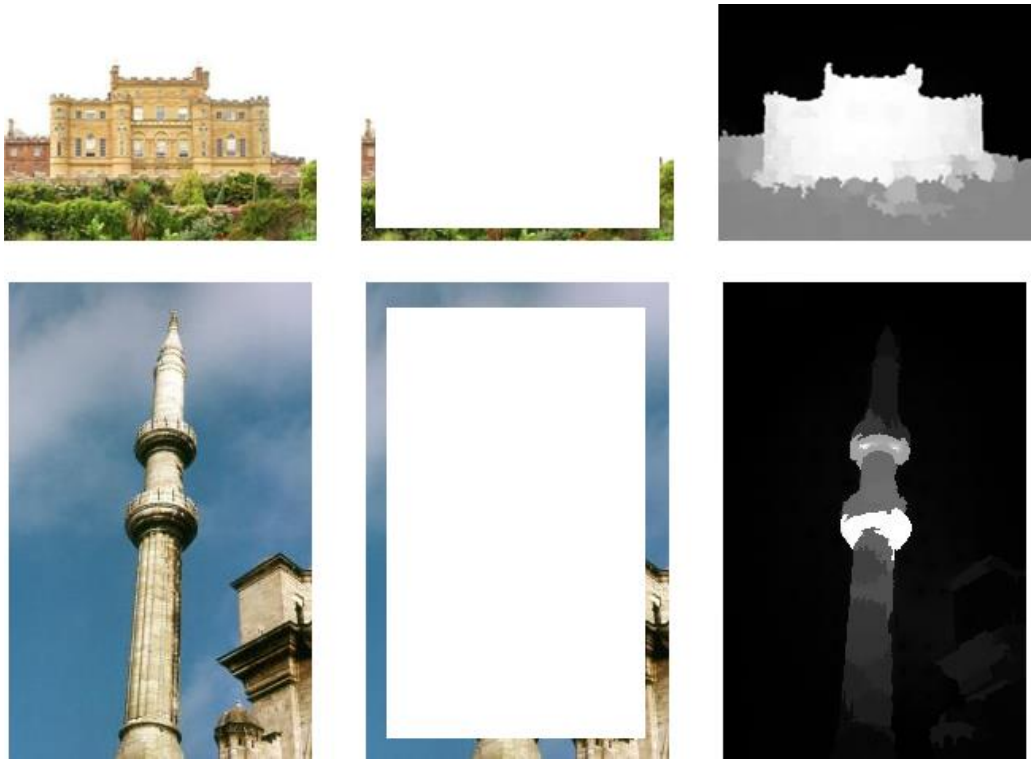


Figure 1.3 Image boundary priors sometimes lead to unsatisfied results. (From left to right: original images, image boundary priors, saliency maps [13] using image boundary priors)

1.2.2. Complex image scenarios

In addition to inappropriate priors, the complex scenarios on natural images are also a big challenge for saliency object detection, especially when precise segmentations of objects are required. For examples, as shown in the first row of Figure 1.4, the saliency detection algorithm is much confused about the case of foreground and background sharing similar appearance in color. Sometimes, as shown in the second row of Figure 1.4, the algorithm cannot cognitively figure out which objects or regions are more attractive to human, when the foreground contains several candidate objects.

Figure 1.4 Complex image scenarios impede the precise saliency object detection. The saliency maps are produced by the method in [13].

1.2.3. Medical image artifacts

Compared to natural images, medical images suffer from more types of artifacts. For prostate MR images, these artifacts include image noise, intensity inhomogeneity and blurred boundaries.

Image noise: Circuit noise [24], transmission noise of imaging equipment [24], inappropriate imaging-testing position of patients and other sources of noise can cause the low quality of images. Gaussian noise, salt-and-pepper noise and speckle noise [25] are the main three types of image noise. The image noise makes the one-channel pixels (image intensities) distorted and less informative, which increases the difficulty of object (anatomical structure) recognition.

Intensity inhomogeneity: Due to the tumors on the tissues of patients, the to-be-tested tissues are usually coarse, which are displayed as inhomogeneous regions in medical images. Such intensity inhomogeneity poses serious challenges for common features learning of the aimed tissues. An example of intensity inhomogeneity on prostate MR image is shown in Figure 1.5.

Blurred boundaries: In clinical practice, some medical images of patients exist blurred tissue boundaries, or even miss the boundaries. This artifact may lead to false-positive results by object recognition algorithms, which decreases the effect of CAD.

Figure 1.5 Inhomogeneous prostate from prostate cancer patient. The prostate region is delineated in red contour.

1.3. Contributions

To address the challenges aforementioned in previous chapters, we propose highly effective and robust methods, with deep networks and image energy, to recognize objects both on natural and biomedical images. The main contributions can be summarized below.

1.3.1. A new comprehensive autoencoder model for prostate recognition

In order to replace user interactions or atlas mappings for prostate seeds selection, we propose a new comprehensive autoencoder model to provide more reliable priors. The contributions of this model include:

1. We propose a new autoencoder-based classifier in which the training set consists of only positive samples, so that it can lessen the impacts by the irregular and complex background that may impede feature extraction.

2. Our proposed model can provide necessary priors for later prostate segmentation and significantly beats classic classifiers on prostate recognition.

3. We extend the model on natural images to recognize salient objects and outperforms some state-of-the-art saliency detection algorithms.

1.3.2. A general object recognition model on multi-modality images

Conventional handcrafted features cannot comprehensively extract intrinsic and latent structures of images for more precise object recognition. To tackle this obstacle, a multi-contexts combined deep neural networks model are proposed in this work. The contribution of this model include:

1. With deep neural networks, the proposed model can semantically and cognitively extract salient objects from complex images and is competitive to most of state-of-the-art methods on popular benchmark datasets.

2. The designed multi-contexts is more adaptive to various images for combination of local and global features, compared to other deep networks based saliency detection methods.

3. The model is also validated on biomedical images (MRI) for prostate recognition, and proved the significant superiority for prior seeds selection.

1.3.3. A novel saliency image energy cooperating with region priors

In order to obtained more saliency maps, we design a graph-cut based image energy for saliency object detection, by imposing region priors on it. The contributions are summarized as:

1. While most graph-cut based energies measure the smooth penalty merely among adjacent pixels, we treat the image as a complete graph in superpixel scale, enabling smooth penalty to be measured in a holistic way.

2. An inherent limitation of complete graph is that it may lead to trivial errors. We therefore used region priors to guide the construction of the smooth penalty.

3. We propose a new saliency object detection method by integrating the proposed saliency image energy and multi-contexts combined deep neural networks model, which outperforms state-of-the-art methods on five benchmarking datasets.

4. Our proposed image energy can adopt any type of saliency map produced by other saliency detection methods, and thus can be a post-process and refinement for most existing approaches.

2. Background

In this chapter, we introduce some related works on object recognition. We first briefly summarize the applications of object recognition both on natural and biomedical images, followed by an introduction of magnetic resonance imaging which we will typically utilize for model validation. Afterwards, we provide the literature review of current object recognition methods and discuss their pros and cons.

2.1. Applications of object recognition

As an important branch of image processing, object recognition can be applied to a wide range of algorithms and scenarios, both on natural image and medical image. Some typical applications of object recognition are as follows:

Facilitation of other image processing tasks: Many object/image classification methods utilize object recognition algorithms to narrow the targets to-be-classified [26, 27]. Some image segmentation methods [28, 29] employ the recognized candidate objects, especially salient regions and objects, as a kind of prior knowledge to boost the segmentation results.

Automatic sophisticated system: The task-driven object recognitions are widely used in computer-integrated vision system, which dramatically improve the life quality and work efficiency. For examples, the digital cameras deploys face detection [30] algorithm to locate the focus; the use of pedestrian detection [31] in car surveillance system can decrease the rate of traffic crash.

Quantification measurement of tissue volumes: With the increasing number of medical images in clinic, computer-aided diagnosis is highly demanded. Automatic object (anatomical structure) recognition on medical images is a part of CAD system, which feeds the further quantification measurements of tissue volumes such as lung nodule classification on CT images [32], lymphoma staging [33] on PET/CT images and brain tumor segmentation on MR images [34].

2.2. Magnetic resonance imaging

Magnetic resonance imaging (MRI) is a non-invasive medical test for disease diagnoses and clinical treatment. Under a strong magnetic field, the protons of to-be-tested body are realigned and spin out of equilibrium when a radiofrequency current is pulsed. After the radiofrequency, the realigned protons will emit various energy according to the type of body organs and tissues; and the MRI sensor can capture such released energy to determine the position of the source, thus is able to estimate the insides of the to-be-tested body. A kind of medicine containing Gadolinium can be applied to the patient intravenously to boost the speed of realignment of protons, which results in a brighter MR image.

Compared with other medical imaging modalities such as TRUS and CT, MRI provides high contrast images for non-bony parts and soft-tissues, and enables the lesion detection and cancer staging [35]. Conducting MRI test is safe to human body, in that it do not use the damaging radiation; for this reason MRI is particularly well suited to frequent imaging for diagnosis and therapy, especially in the brain. For prostate test, MRI produces a set of tomographic slices of a prostate volume. As shown in Figure 2.1, in addition to prostate, other tissues and organs (i.e. bladder, hip and rectum) near the prostate are also displayed on the image.

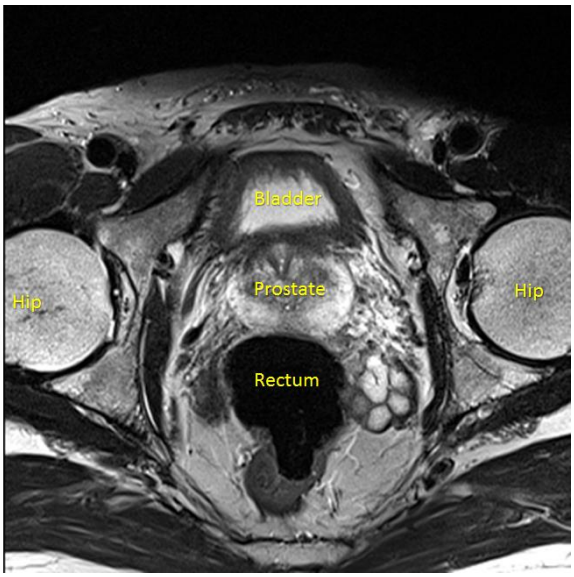


Figure 2.1 A typical prostate and nearby tissues and organs on MR image.

2.3. Literature review on object recognition

2.3.1. Saliency object detection

Bottom-up approaches: Bottom-up based saliency object detection is a data-driven task composed of two stages: feature extraction and saliency computation. Numerous low-level stimulus, such as color, texture and oriented filter responses, have been developed or employed as features. Following the feature extraction, the saliency map can be estimated at single or multiple scales by random walk [36], manifold ranking [14], cellular automata [12] etc.

Since graph model [36] has been first introduced to saliency object detection, saliency propagation is gaining much popularity in recent years. Based on a constructed directed/undirected graph, saliency propagation is to propagate saliency values from labelled pixels (prior seeds) to unlabelled pixels. The popular propagation formulas may include random walk [36] and personalized PageRank [37]. However, in addition to the adopted propagation formulas, inappropriate prior

seeds and weights of graph edges also greatly affect the accuracy of results. In order to address such issues, Li et al. [13] pre-processed the original image with the use of random walk to trim the set of prior seeds. Inspired by the common teaching mechanism in real-world classes, Gong et al. [38] proposed a progressive propagation which predicts saliency values from 'simple' image patches to 'difficult' image patches.

Top-down approaches: Top-down based saliency object detection is task-driven which models the binary classification (i.e. background group and foreground group) via a set of training images. Compared with bottom-up approaches, top-down methods do not highly rely on the prior seeds and weights of graph, yet still requires carefully feature extraction. Gao et al. [39] pre-defined a filter bank to extract discriminant features which are dominated by the target regions in training set. Instead of the filter bank, Liu et al. [40] computed saliency values via Conditional Random Field (CRF) which is a flexible framework for feature incorporation in saliency detection. As an improvement of CRF learning, Yang et al. [41] proposed CRF supervised sparse coding to learn the saliency computation model including CRF weights and sparse coding dictionary.

Deep learning approaches: The high-level features extracted from deep neural networks (DNN) can lead to a promising results in saliency detection, beating the most conventional methods with a significant gap. In the recent two years, one of the most popular DNN adopted in saliency detection is convolutional neural networks (CNN) [42] which emulates the functions in the animal visual cortex.

According to the size of operated units, CNN based saliency detection can be categorized into pixel-wised, superpixel-wised and region-wised methods. Pixel-wised approaches [43] generally take the whole image as input and directly output

the pixel-wised saliency map from a very deep neural network, such as fully convolutional networks (FCN) [44]. Superpixel-wised approaches [45] conduct the algorithms superpixel by superpixel, and then merge the estimations of each superpixel as the final saliency map. Region-wised approaches [46] usually exploit efficient image segmentation algorithms (e.g. globalized probability of boundary based contour detection [47]) to first partition the image into several sub-regions, and then extract the high-level features of each sub-regions by CNN. The experiments of these recent deep learning based methods demonstrate that the high-level features can depict the latent and intrinsic structures of input data, while the low-level cues do not have such capacity.

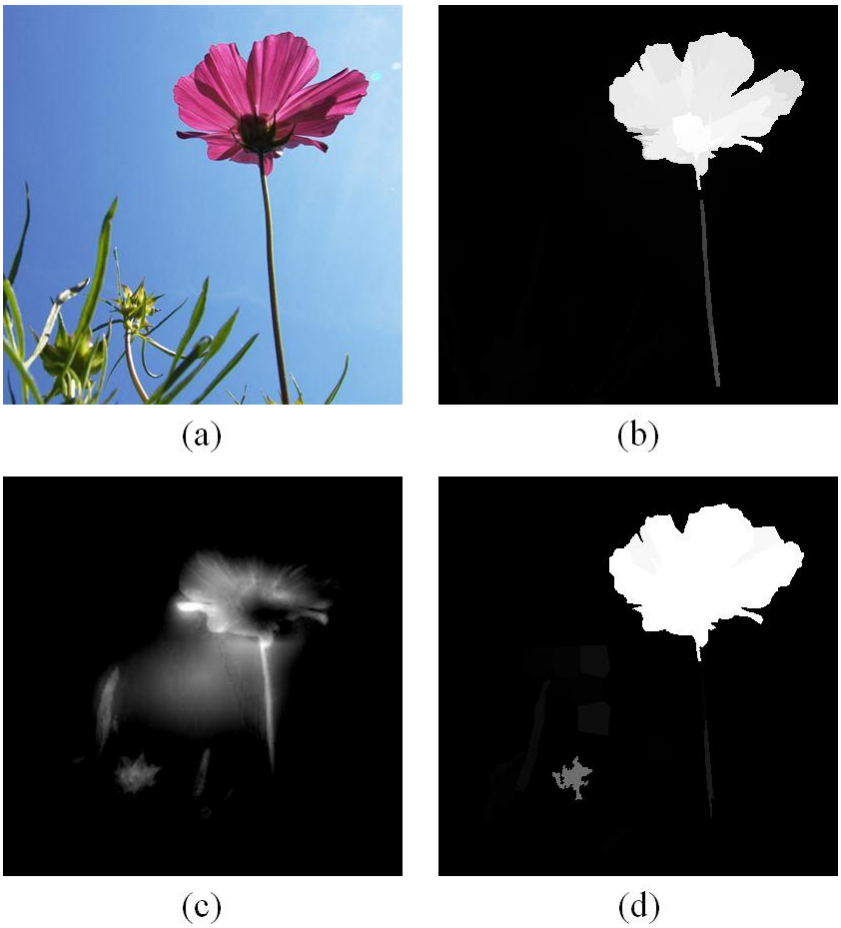


Figure 2.2 Saliency maps of input image (a) generated by: (b) bottom-up approach [14], (c) top-down approach [48], and (d) deep learning approach [45].

2.3.2. Prostate recognition

Atlas mapping: The works on anatomical structure recognition on medical images are rare, which makes the further tissue segmentation and cancer qualification tough to perform. Many anatomical structure segmentation approaches, e.g., [49, 50] are often limited by the recognition techniques in medical imaging, as accurate segmentation often requires approximate localization of the target anatomical structure as initialization. To address this challenge, conventional segmentations on medical images rely on semi-automatic methods thereby being dependent on the user [51-53].

Alternative approaches explore the use of an image atlas to define the foreground/background prior seeds [49, 50]. The atlas is a global probabilistic cloud, respective to a specific type of imaging, such as prostate MR images [49] and liver CT images [54]. By stacking a set of human annotations (binary maps as shown in the second column of Figure 1.2), the density of each pixel/voxel on atlas indicates the corresponding likelihood the pixel/voxel being foreground. The atlas is then registered with a specific testing image so that it is applicable to the testing image. With the registered atlas, the foreground seeds can be selected by a defined threshold. However, as noted in prior studies [35], reliance on atlas are still prone to generating errors.

Contour and shape based approaches: A set of works, such as contour-based methods and deformable model based methods, exploit contour and shape information for prostate segmentation. Contour-based methods [51-53] usually extract edges and ridges in images via gradient filters, and recognize or trace the boundaries by their proposed schemes (e.g. the longest curvi-linear structure [52] and

moving masks [51, 53]). However, the edge detectors may not be always reliable due to the artifacts (blurred/broken boundaries) on biomedical images.

Since deformable model was first introduced by Terzopoulos [55], it has been widely applied in many prostate segmentation works [56-62] with the utilization of contour and shape information. Deformable models are curves or surfaces and usually formed under the control of internal and external energies [63]. Internal energies preserve the smoothness of curves (surfaces) during deformation, and external energies force curves (surfaces) towards the anatomical structure boundaries [35]. By minimizing the joint internal and external energies, the deformable models can be evolved to the desired positions. Active shape model (ASM) [64] is one of the most popular modalities used in deformable prostate models [60, 65-70]. In ASM based methods, a statistical shape model (SSM) [71] is constructed with shape variations using principle component analysis (PCA) on a set of landmarks, and then ASM is performed to delineate the target objects. As ASM overlooks the interdependencies of shape and appearance [72], active appearance model (AAM) [73] thus is developed for the purpose of combination of shape and appearance. However, as noted in [72], conventional ASM and AAM based methods are hindered by the use of landmarks. To solve this issue, Toth et al. [72] proposed a novel landmark-free AAM based methods for more accurate and robust prostate segmentation on MRI. Other modalities applied in deformable model for prostate segmentation include level set [74-79], active contour model [80-85] and so on.

Graph based approaches: Many prostate segmentation works [86-94] transform prostate images to (un)directed graphs, usually followed by a cost function. The atomic units (pixels, voxels, superpixels or supervoxels) are the nodes of graph, and the edge weights are represented by the ‘distances’ of pairs of nodes. The

essential parts of these graph based approaches are the design of edge weights on graphs and cost functions. Positions [86] and intensities [86, 90, 94] of pixels (voxels, superpixels or supervoxels) are the two extensively used measurements for edge weights. As the utilization of position and intensity are limited by morbid biomedical images with low contrast or distorted prostate, some works employ other information, such as prior shape knowledge [91] and image gradients [93], to estimate edge weights. The cost functions are various across the graph based methods, but most of them [87, 88, 90, 92, 94] are formulated from graph-cut model [95]. In addition to graph-cut, Lagrange function [86] and other special designed functions (e.g. shape probability function and gradient profile model in [93]) can also be applied to energy minimization scheme. However, the fixed parameters for balancing cost function need carefully tuning so that may hinder the robustness of methods across different datasets.

After the construction of graph model, some other works [49, 96, 97] formulate segmentation as labelling propagation problem, in which unlabeled nodes can be predicted by pre-defined labelled nodes. Random walker [98] has been proved an effective and efficient algorithm to solve the labelling propagation problem in prostate segmentation [49, 96, 97]. As such propagation requires foreground/background seeds, this kind of graph-based approaches [96, 97] are usually semi-automated with user interactions. More recently, by employing atlas mappings as priors, a fully automated prostate segmentation algorithm with enhanced random walker is proposed in [49], however still gets trapped into the wrong seeds produced by atlas.

Classification based approaches: The classification based approaches extract a set of image features as feature vectors, and tend to partition feature space

(vector space associated with feature vectors [35]) into two or more groups. The classic classifiers, such as support vector machine [99] and random forest [100], have been extensively studied in the last decades and proved favorable capacity of feature space partition, thus can also be applied in prostate segmentation works. Gray level intensity and spatial coordinate are the simple but useful common features that are widely used in many works [101-104]. Other computer vision features, such as histogram of oriented gradients [105, 106], Haar features [105, 106], curvature [103], Haralick texture features [104] and Laws energy features [104], are also widely employed to differentiate prostate. Ghosh et al. [107] imposed prior knowledge on texture and shape features by genetic algorithm, which achieved better segmentation results compared to Laws energy features. Instead of classic classifiers, Li et al. [106] proposed a set of location-adaptive classifiers which enable to effectively gather local information and propagate them to other regions. In the work of [105], Gao et al. proposed an extended sparse representation based classification to address the issue of low contrast on prostate images. Although the aforementioned features technically enables to differentiate prostate, such low-level descriptors cannot extract intrinsic structures of images and thus are still insufficient for more precise prostate recognition and segmentation.

Hybrid segmentation: As the hybrid of techniques are robust to noise and produce superior results in the presence of shape and texture variations of the prostate [35], most works combine two or even more methods for prostate segmentation.

As a common prostate segmentation approach, deformable models are usually combined with various techniques to boost the performance. Graph based methods and classification based methods are usually employed to initialize

deformable models in many works [56, 59, 76, 108, 109]. Zhan et al. [59] tentatively labelled voxels by proposed Gabor-SVM classifier to feed the later deformable surface model. More straightly, Martin et al. [56] utilized atlas to map a specific prostate image before deformable model. For more reliable initialization of deformable model, Guo et al. [108] employed deep learning features to estimate rough prostate recognition map. Different from aforementioned works exploiting priors for deformable models, the results by deformable models in the work of [110, 111] can also be treated as location and shape priors for other techniques (i.e. Bayesian classification).

Classification based approaches are usually followed by a graph-cut based cost function in a set of works [87, 90, 103, 112, 113]. On one hand, such combination focuses on local features in patch classification phase; on the other hand, the correlations of neighboring patches/pixels can be taken into account for smoothness in cost function. Other hybrid segmentation methods can be found in [50] (atlas and shape model combined), [114] (level set and registration combined) and [115] (representation learning and labelling propagation combined).

3. A new comprehensive autoencoder model for prostate recognition

Automated anatomical structure recognition is an essential prerequisite in precision medicine such as tissue segmentation, physiological signal measurement and disease classification. It poses a challenging task because of the insufficient color information of pixels and low signal-to-noise ratio in medical images [51]. Previous works have been proposed to tackle anatomical structure recognition problems based on handcrafted features, such as steerable feature, on a wide array of imaging modalities, e.g., ileocecal valves [116], polyps [117], and livers [118] in abdominal CT, and heart chambers in ultrasound [119]. However, to our best knowledge, no work has been done on prostate recognition in MR images, although prostate cancer accounts for the second highest mortality rate among various types of cancer on males [120] and MR images prove effective for prostate diagnoses and treatments [35]. In addition to the insufficient color and speckle information, MR image artifacts, such as low contrast and blurred tissue boundary, make it even more difficult to accurately locate the prostate.

In this chapter, we propose a novel prostate recognition method on MR images which combines handcrafted features with deep autoencoder networks. Autoencoder (AE) is an unsupervised learning algorithm and is capable of extracting and reproducing the statistical structure for a given dataset [121]. Different from the most works which embed a classifier on the top of the last layer in deep neural network [11, 122], we propose a novel method to compute prostate recognition map through taking advantage of outstanding capability of autoencoder for data

reconstruction. Afterwards, we design an image energy minimization scheme to generate a stronger prostate recognition map with consideration of the relationship among neighboring pixels. The following methods are based on our previous works in [2].

3.1. Prostate recognition method

As shown in Figure 3.1, our prostate recognition method consists of four stages. Firstly, early feature descriptors are extracted to feed the proposed stacked autoencoder. Secondly, we train a prostate stacked autoencoder (SAE) classifier in iteration. Thirdly, the likelihood of a pixel belonging to the prostate can be estimated via our proposed new algorithm. Lastly, an image energy minimization scheme is applied to optimize the recognition result.

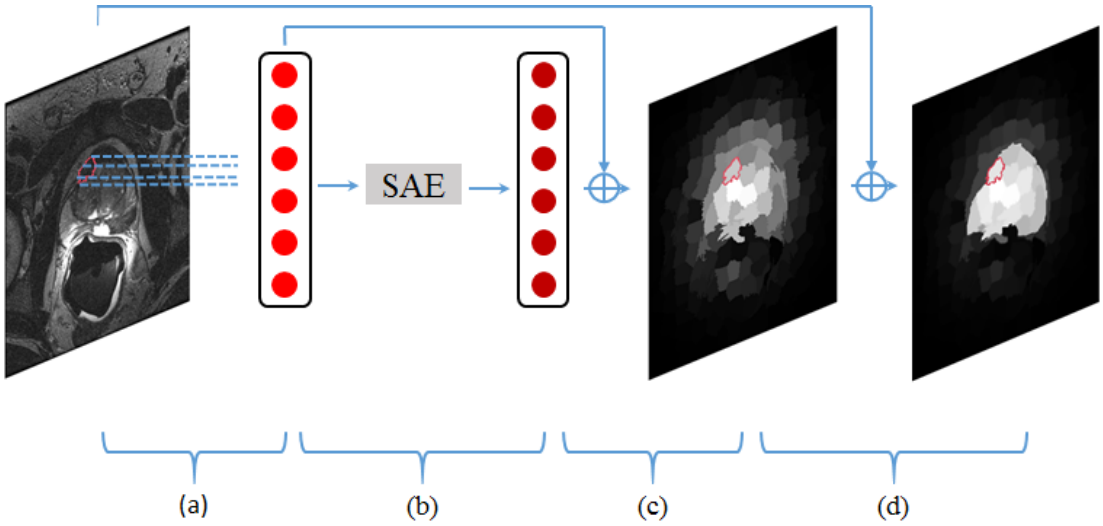


Figure 3.1 Pipeline of our method. (a) Early feature extraction. (b) Superpixel reconstruction via proposed prostate AE model. (c) Superpixel classification. (d) Refinement via proposed image energy.

3.1.1. Early feature descriptors

Instead of merely using the pixel intensity values, we adopt two early features, i.e. the intensity descriptor and the position descriptor as the input for the deep autoencoder network, which reflects pixel-value and spatial information respectively. Formally, an image $I \in \mathbb{R}^{m \times n}$ is segmented into N superpixels via the SLIC algorithm [123]. We denote a superpixel as P. As suggested in [124], the superpixel is first whitened via zero phase component analysis (ZCA) to make the pixels less correlated with each other. An early feature vector $f(P)$ is then derived for P with details as follows.

Intensity descriptor: Intensity histogram is an effective measure to describe the intensity distribution of an image patch. Hence, we adopt the intensity histogram $IH(P)$ as the intensity descriptor for superpixel P. In our experiment, the number of bins is set to 20 empirically. Then, the intensity histogram $IH(P) \in \mathbb{R}^{20 \times 1}$ is normalized to have a uniform sum to eliminate the effect caused by the different number of pixels within different superpixels.

Position descriptor: From our observation, most prostate tissues are approximately located at the centre area of patient MR image. This is the assumption on which many works are based, especially those where probabilistic atlases were employed [49, 50]. Thus, such prior knowledge is informative for prostate detection. Since the superpixels are of irregular shapes, we exploit bounding boxes to approximate their spatial locations. We denote the bounding box of P as

$$C(P) = \{c_v(\alpha_{v,1}, \alpha_{v,2}) : v = 1,2\} \tag{3.1}$$

where c_1 and c_2 are the top-left coordinate and bottom-right coordinate of $C(P)$ in image $I \in \mathbb{R}^{m \times n}$ respectively. $\alpha_{v,1}$ and $\alpha_{v,2}$ are c_v's values corresponding to x-axis

and y-axis respectively. The position descriptor $POS(P) \in \mathbb{R}^{4\times1}$ of superpixel P is then calculated by

$$POS(P) = \left\{ t(v,u) = \frac{\alpha_{v,u}}{(2-u)n + (u-1)m} : v = 1,2; u = 1,2 \right\} \tag{3.2}$$

Early feature vector: With the early feature descriptors proposed above, a superpixel-wise feature vector $f(P)$ with 24 dimensions is generated as

$$f(P) = \{IH(P); POS(P)\} \in \mathbb{R}^{24\times1} \tag{3.3}$$

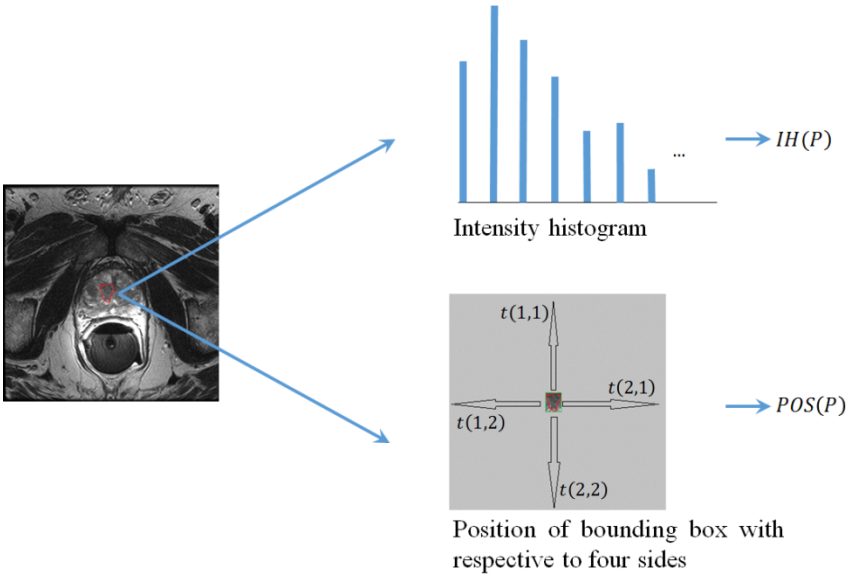


Figure 3.2 Illustration of Intensity descriptor and position descriptor.

3.1.2. Prostate stacked autoencoder model

After obtaining the early feature vectors of prostate superpixels, we can build a stacked auto-encoder (SAE) to extract high-level features and perform the reconstruction of input early feature vectors for later classification. An autoencoder consists of encoding process and decoding process. In the encoding process, the AE tends to learn a set of encoding weights to construct a code vector given the input

vector; similarly, in the decoding process, it learns another set of decoding weights to map the code vector into an approximate reconstruction for the input vector.

To train a single-hidden-layered prostate AE, a training set $F = \{f(P_i): i = 1,2,\dots,K\}$ containing K early feature vectors of prostate superpixels are input to the AE. Each node is fully connected by undirected weight matrix with an associated bias value between each layer (i.e. input layer, hidden layer and output layer). The input vector $f(P_i)$ is transformed into a hidden feature representation a_i by an activation function $g(\cdot)$ with the following formula:

$$a_i\left(f(P_i);\theta^{(1)}\right) = g(W^{(1)}f(P_i) + b^{(1)}) \tag{3.4}$$

where $\theta^{(1)}$ is the parameter vector including weight matrix $W^{(1)}$ and bias term $b^{(1)}$; as a common practice, we use the sigmoid function $g(\phi) = 1/(1+\exp(-\phi))$ as the activation function. A decoder then maps the hidden feature representation a_i back to an approximate reconstruction $\hat{f}(P_1) \in \mathbb{R}^{24\times 1}$ in a similar transformation

$$\hat{f}(P_1)\left(a_i;\theta^{(2)}\right) = g(W^{(2)}a_i + b^{(2)}) \tag{3.5}$$

With the training set F of K samples, the latent features of input data can be learned by minimizing the cost function

$$J(\theta) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{2}\left\|f(P_i) - \hat{f}(P_1)\right\|^2 + \frac{\lambda}{2}\sum_{i=1}^{s^{(1)}}\sum_{j=1}^{s^{(2)}}(W_{ij}^{(1)})^2 \tag{3.6}$$

where the first term in $J(\theta)$ is an average sum-of-squares error term and the second term is a weight decay term that tends to decrease the magnitude of the weight and prevent overfitting [125], with a weight decay parameter λ. $s^{(1)}$ and $s^{(2)}$ are the numbers of nodes in the first layer (input layer) and second layer (hidden layer) respectively. A sparsity constraint is usually imposed on the hidden nodes to enhance

the probability of linear separability [126] and the overall cost function (5) is modified as

$$J(\theta) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{2}\left\|f(P_i) - \hat{f}(P_1)\right\|^2 + \frac{\lambda}{2}\sum_{i=1}^{s^{(1)}}\sum_{j=1}^{s^{(2)}}(W_{ij}^{(1)})^2 + \beta\sum_{j=1}^{s^{(2)}}KL(\rho||\hat{\rho}) \quad (3.7)$$

$$KL(\rho||\hat{\rho}) = \rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j} \quad (3.8)$$

where ρ is a small value close to zero, which specifies the desired level of sparsity. $\hat{\rho}_j = \sum_{i=1}^{K}[a_i]_j/K$ is the average activation of the j-th hidden node and the Kullback-Leibler (KL) divergence provides the sparsity constraint. β is the weight of the sparsity penalty term. We use gradient descent optimization algorithm [127] to update θ in iterations and back-propagation algorithm is applied to calculate the partial derivatives in this process.

As [126, 128] suggested, to fully utilize the ability of deep neural networks, we further construct stacked autoencoder (SAE) to perform feature presentation to learn highly nonlinear and complex patterns in the input images. As shown in Figure 3.3, in a stacked autoencoder structure, the original data, i.e. the early feature vector, is input to the first (bottom) auto-encoder, and its hidden nodes (or units) are concatenated as a new feature vector which is used as the input data for training the subsequent (higher-level) auto-encoder. The greedy layer-wise algorithm is adopted to obtain the corresponding parameter $\theta^{(l)}$ of the l-th layer. After the training of each sub-AE is complete, back-propagation is applied again to tune the parameters of all layers at one time. Typically in our work, we stack three AEs to construct the prostate SAE model and hence obtain a totally six layer network including three encoding layers and three decoding layers.

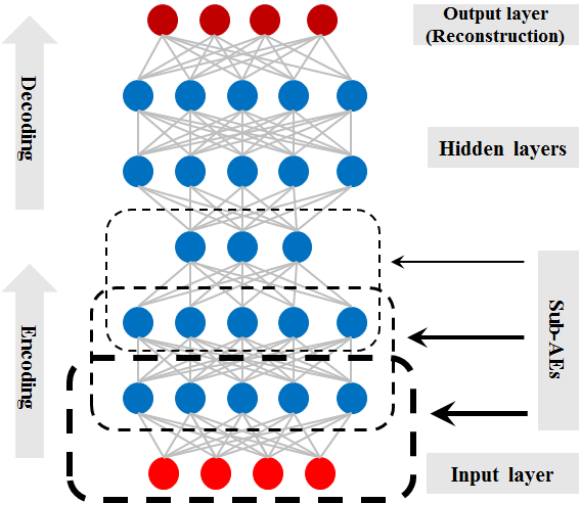


Figure 3.3 Architecture of the proposed SAE. The output of each layer is the input for its subsequent layer. The output of the last layer is a reconstruction for input data.

3.1.3. Superpixel classification

With the well trained SAE model, the superpixels of the input image can thus be classified. Different from other deep learning algorithms (i.e. convolution neural network), not only can auto-encoder learn intrinsic and latent feature presentation for input data, it is also capable of data reconstruction. Therefore, we can calculate the reconstruction errors for each superpixel in a prostate MR image via the fixed prostate SAE model. Specifically, with the all parameters $I = \{\theta^{(l)}: l = 1,2, \dots, L\}$ learned in SAE, for a superpixel P, set

$$f(P)^{(l+1)} = g(W^{(l)} f(P)^{(l)} + b^{(l)}) \tag{3.9}$$

where l is the index of network layer. We initialize the first step of the iteration $f(P)^{(l)}$ as the early feature vector $f(P)$ of the superpixel P. Then the reconstruction error of P is calculated by $f(P)$ and $f(P)^{(L+1)}$:

$$err(P) = \sum_{\omega=1}^{24} \exp(\tau \| f(P)_{\omega} - f(P)^{(L+1)}{}_{\omega} \|^2) \tag{3.10}$$

where $f(P)_\omega$ and $f(P)^{(L+1)}{}_\omega$ are the ωth elements of $f(P)$ and $f(P)^{(L+1)}$ respectively. τ controls the distance between different superpixel's reconstruction errors within an image and is set to 100 empirically. We adopt the reconstruction error to measure the probability of a superpixel being prostate tissue. This is because as the SAE model is learned from the set of prostate superpixels, the prostate superpixel should have a lower reconstruction error than the background superpixel does and vice versa.

After calculating the reconstruction errors of all the superpixels in an image $I \in \mathbb{R}^{m\times n}$, we may obtain a weak prostate detection map $D^{AE} = \{d^{AE}{}_i \in [0,1]: i = 1, \dots, m \times n\}$. D^{AE} is calculated without considering the spatial and intensity coherence among superpixels, hence it is a local estimation. In the next sub-section, a refined prostate detection map with better supressed background, more smooth inner region and clear boundary is generated based on D^{AE}.

3.1.4. Refinement

Given a one-channel image I, our task in this stage is to assign a label $O_p \in \{0,1\}$ to a pixel p to measure whether p belongs to foreground or not. For the set of pixels' labelling $O = \{O_p: p \in I\}$, this can be solved by minimizing the energy function [15]

$$E(O) = \sum_{p\in I} H(O_p) + \xi \sum_{(p,q)\in Y} \frac{1}{1+\sqrt{3(I_p - I_q)^2}} \cdot T(O_p \neq O_q) \tag{3.11}$$

where Y is a set of all pairs of neighboring pixels. $H(O_p)$ is the cost for assigning a label O_p to a pixel p. We directly use local estimated detection map D^{AE} to approximate the label-cost of pixels. Specifically, $H(O_p)$ is set to $D^{AE}(p)$ if O_p is a background label and $1 - D^{AE}(p)$ if O_p is a foreground label. The second term in

(3.11) encourages intensity and spatial coherence by penalizing discontinuities [95] between neighboring pixels, with the parameter ξ controlling the scale of discontinuity penalty. $T(\cdot)$ is 1 if the condition inside the parentheses is true and 0 otherwise.

We adopt minimum cut/maximum flow algorithms [95] to minimize (3.11) and generate the corresponding prostate detection map D^{mf}. Then D^{AE} and D^{mf} are linearly combined as the final prostate detection map:

$$D = \frac{D^{AE} + D^{mf}}{2} \tag{3.12}$$

We directly use D to measure the probability of each pixel being prostate.

3.2. Experiment and evaluation

3.2.1. Setup and dataset

The prostate MR Image Segmentation 2012 (PROMISE12) database [129] is used in this study. It contains 50 cases, with each case composed of 15 to 54 prostate transverse T2-weighted MR images. Manual segmentation are available for each case and used as the ground truth.

In the prostate SAE model, the hyperparameters of each sub-AE, i.e. the number of hidden nodes Z, and weight decay parameter λ, are derived empirically and listed in Table 3.1.

Table 3.1 Hyperparameters in the prostate SAE model.

	sub-AE 1	sub-AE 2	sub-AE 3
Z	60	40	16
λ	8e-4	4e-4	4e-4

3.2.2. Evaluation metrics

Following the works of [8, 13, 15], we adopt precision-recall (PR) curve, F-measure and mean absolute error (MAE) to evaluate the performance of our proposed method. Specifically, precision and recall are defined as

$$precision = \frac{\sum_{i \in A} A(i) \cdot B(i)}{B(i)} \tag{3.13}$$

$$recall = \frac{\sum_{i \in A} A(i) \cdot B(i)}{A(i)} \tag{3.14}$$

where A and B are the ground truth and saliency map by the algorithm respectively and both normalized in the range of [0, 255]. Then, we binarize the continuous saliency map with the fixed threshold from 0 to 255 with an increment of 1 to construct the PR curve.

Generally, neither $precision$ and $recall$ can individually and comprehensively evaluate a certain algorithm [130]. For this reason, a harmonic metric (i.e. F-measure) is adopted to measure the comprehensive performance of an algorithm:

$$F_\eta = \frac{(1 + \eta^2) \cdot precision \cdot recall}{\eta^2 \cdot precision + recall} \tag{3.15}$$

where η is to balance the weights of $precision$ and $recall$. As high $recall$ can be easier achieved compared to high $precision$ (e.g. simply full foreground map leads to 100% recall score), η^2 is usually set to 0.3 to emphasize the weight of $precision$ [8, 13, 14, 130].

As PR curve and F-measure focus on the true positive saliency assignments, i.e. recognizing salient region, we adopt MAE score to measure the results of non-saliency recognition by a certain algorithm. MAE is defined as

$$MAE = \frac{\sum_{i \epsilon \bar{A}} |\bar{A}(i) - \bar{B}(i)|}{N_{pixel}} \tag{3.16}$$

where $\bar{A}$ and $\bar{B}$ are the ground truth and saliency map by the algorithm respectively and both normalized in the range of [0, 1]; N_{pixel} is the number of pixels on the image. A lower MAE score means the better capacity of minimizing the gaps between ground truth and saliency map.

3.2.3. Experimental results

As suggested in [15], to achieve better performances, we computed five recognition maps using five superpixel scales with $N = 200, 250, 300, 350, 400$ respectively in an image. Then, we linearly combined the five recognition maps as the final recognition result. For each image, we resized it to 320∗320 pixels, and increased its contrast by mapping the intensity values to new values such that 1% of data is saturated at low and high intensities of the image [131]. 10-fold cross validation was performed here on the PROMISE12 dataset. As shown in Figure 3.4, our proposed stacked autoencoder can recognize the position and rough shape of the prostates. After the image minimization, the obtained recognition maps are more accurate and even can segment the aimed prostate.

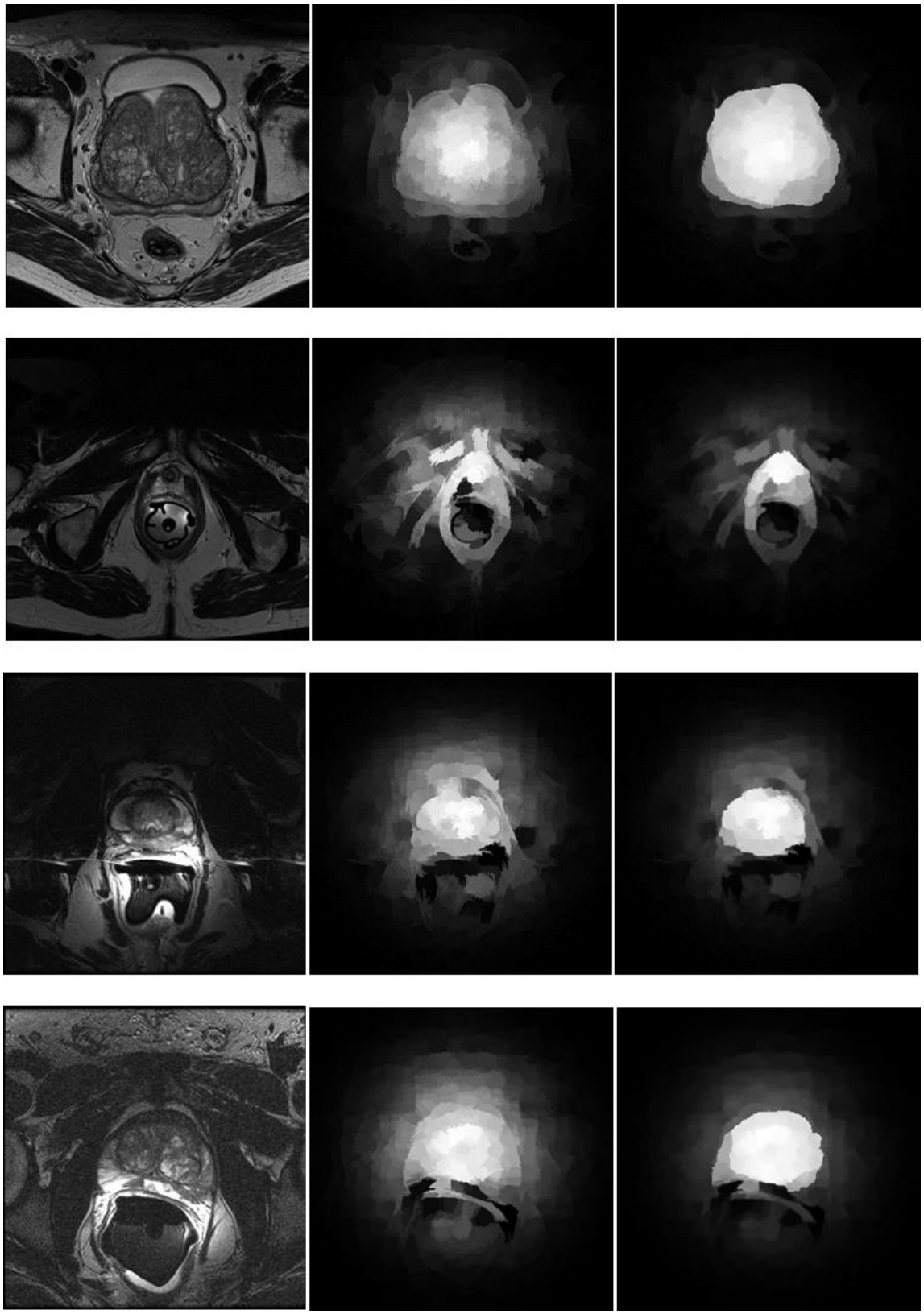


Figure 3.4 Examples of prostate recognition results by our method. Left to right: original prostate MR image, rough recognition map by proposed prostate stacked autoencoder, and final recognition map.

3.2.4. Evaluation

We evaluated the recognition performance using precision-recall (PR) curve and F-measure [15, 132]. An atlas-based seeds-selection in segmentation approach (RW) [49] and three popular classifiers, i.e. support vector machine (SVM) with radial basis function kernel, random forest (RF), and naive Bayes (NB), were chosen as comparison methods.

Both Table 3.2 and Figure 3.5 demonstrate that our method outperform the four comparison methods in terms of both PR curve and F-measure. More specifically, even our unrefined results outperform the refined results of the comparison methods in precision. This is mainly attributed to the SAE for high-level feature learning and data reconstruction, while the comparison methods recognize prostate directly from the low-level early features. Figure 3.4 qualitatively demonstrates that our proposed refinement significantly contributes to foreground smoothness and background suppression. The refinement poses relatively low effect around the prostate with blurred boundary as illustrated in the second row of Figure 3.4. The reason is that the neighboring pixels around the boundary does not differentiate much, thus causing a large penalty in the second term of (3.11), which encourages to assign same labels to these pixels around the boundary of prostate. However, from Table 3.2, it can be seen that our proposed refinement improves all the methods in precision and F-measure.

Table 3.2 Precision and F-measures of our method and comparison methods for prostate recognition on PROMISE12 database, and the Pearson product-moment correlation coefficient (PPMCC) of the two steps. The best results in each column are shown in bold.

	Precision		F-measure	
	Not refined	Refined	Not refined	Refined
OURS	**0.8515**	**0.8699**	**0.6798**	**0.6832**
RW	0.8284	0.8286	0.6617	0.6220
SVM	0.5554	0.6394	0.5415	0.6238
RF	0.4870	0.5506	0.5189	0.5766
NB	0.3539	0.4894	0.3906	0.5033
PPMCC	0.9943		0.9258	

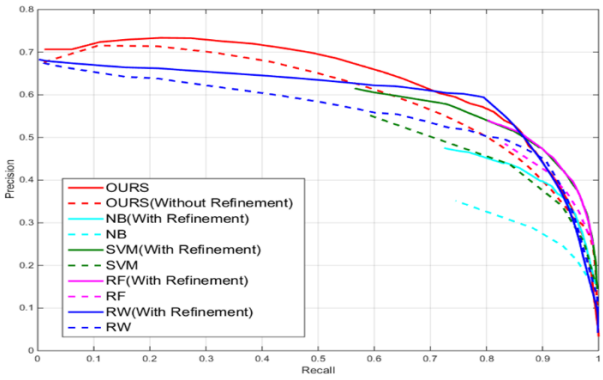


Figure 3.5 PR curves of our method and comparison methods for prostate recognition on PROMISE12 database. The recognition results by comparison methods are also refined by our proposed approach for better evaluation (solid lines).

3.3. Application of saliency object detection

We have proposed an automatic prostate recognition method on MR images based on SAE. One of the major contributions is that we let the SAE itself serve as a

classifier to focus on the prostate feature extraction. Inspired by this idea, in this chapter, we try to extend the prostate stacked autoencoder model to recognize salient objects on natural image. The works of [20, 126] have studied the AE in saliency detection. However, [126] focused on saliency fixation prediction and cannot be directly applied in saliency object detection. In [20], they only utilized AE for classification and still heavily relied on boundary-background priors.

3.3.1. Method extension

In order to obtain reliable prior seeds, we first propose an AE-based approach to search the background seeds. Afterwards, another AE based on prostate stacked autoencoder model is performed hierarchically to form the final saliency map via data reconstruction capability inherent in AE.

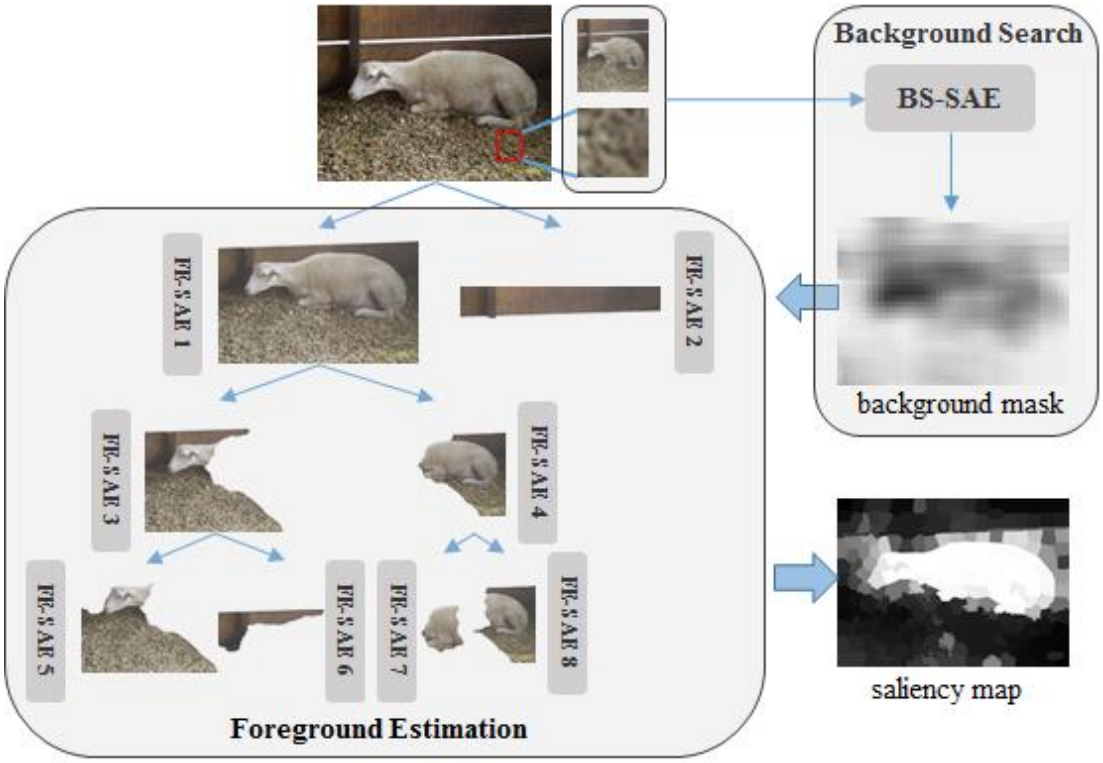


Figure 3.6 Overview of the extended method for saliency object detection.

Background search: For a three-channel image patch p_{bs} with the size of $m \times m$ pixels from the training image I, the input vector $f(p_{bs})$ of background search SAE (BS-SAE) is obtained by

$$f(p_{bs}) = \begin{bmatrix} g(p_{bs}) \\ g(\check{I}) \end{bmatrix} \tag{3.17}$$

where $\check{I} \in \mathbb{R}^{m \times m \times 3}$ is the resized image of I, and following [8], m is set to 51 in this work; $g(\cdot)$ is the vectorization operation, and thus we have $f(p_{bs}) \in \mathbb{R}^{15606 \times 1}$. With the feature representations of each image patch by the trained BS-SAE model, we use softmax regression to measure the probability of each image patch being background. This generates a background mask M_{bs} of I, which can be utilized for further foreground estimation. As shown in Figure 3.7, compared to the conventional boundary-background priors [13-15, 17, 20, 133, 134], such background mask can capture the background region semantically and cognitively, thus it is adaptive for background search.

Figure 3.7 Examples of background mask by BS-SAE.

Foreground estimation: We then extend the prostate stacked autoencoder model for estimation of finer object saliency, with the guidance of the background mask. To improve the efficiency of our algorithm, we transform M_{bs} to a superpixel-wise background mask and use superpixel as the atomic unit in further operation. This can be easily implemented by calculating the mean value of pixels belonging to one superpixel as the probability of the superpixel being background. For brevity, we use M_{bs} to denote the superpixel-wise background mask unless otherwise specified.

With the testing image I and the corresponding background mask M_{bs}, we construct the foreground estimation SAE model (FE-SAE) to extract the foreground of I. Different from the BS-SAE model, the RGB histogram of the superpixel, with 20 bins in each color channel, are exploited as the input vector of the FE-SAE; and there is no softmax regression in FE-SAE, thus it is totally an unsupervised learning

model. Only those superpixels whose values on M_{bs} are more than 0.7 are selected as the training set for the FE-SAE model.

After the training of FE-SAE, we calculate the reconstruction residual $r_{p_{fe}}$ for each superpixel p_{fe} of I by

$$r_{p_{fe}} = \left\| h(p_{fe}) - \bar{h}(p_{fe}) \right\| \tag{3.18}$$

where $h(p_{fe})$ is the original input vector corresponding to p_{fe} and $\bar{h}(p_{fe})$ is the data reconstruction of $h(p_{fe})$ by FE-SAE. Following the idea of proposed prostate stacked autoencoder model, as the FE-SAE is constructed by the background superpixels, the superpixels belonging to background have low reconstruction residual, while those belonging to foreground have high reconstruction residual. Hence, we use the reconstruction residual to measure the saliency value of p_{fe} with the following formula:

$$\begin{cases} s_{p_{fe}} = \dfrac{1}{1 + e^{\frac{\xi\left(u - r_{p_{fe}}\right)}{u - v}}} \\ u = \max\{r_p : p \in \mathcal{D}\} \\ v = \dfrac{1}{|\mathcal{D}|} \sum\limits_{p \in \mathcal{D}} r_p \end{cases} \tag{3.19}$$

where ξ is the smooth factor and set to 6 empirically; r_p is the reconstruction residual of superpixel p by (3.18); and $\mathcal{D}$ is the training set of FE-SAE.

Considering the complex background which may impede the precise foreground estimation, we hierarchically conduct foreground estimation algorithm in regional scales for better performance. Specifically, the testing image I is first segmented into two regions by Ncut algorithm [135]. Two individual FE-SAEs are then constructed respectively under the two regions and each superpixel of I is

assigned to the saliency value by (3.19) with the corresponding FE-SAE. In the next hierarchy, we segment the two regions respectively to generate four smaller regions and construct four individual FE-SAEs corresponding to these regions. Each superpixel of I is assigned to the new saliency value by (3.19) in this hierarchy. Note that in each segmentation operation, only two sub-regions are generated and the region is no longer segmented when $|\mathcal{D}'| \leq 0.3 \times |\mathcal{A}|$ or $|\mathcal{D}'| \geq 0.7 \times |\mathcal{A}|$, where $\mathcal{D}'$ and $\mathcal{A}$ are the training set and superpixel set respectively corresponding to the region. This process is repeated until there regions to be segmented are exhausted. Finally, the saliency value of the superpixel is obtained by linearly combining the saliency values of each hierarchy. The constructed binary segmentation tree is shown in Figure 3.6 and the hierarchical foreground estimation algorithm is summarized in Algorithm 1.

Algorithm 1: Hierarchical Foreground Estimation

Input: testing image I, background mask M_{bs}
Output: saliency map $S = \{s_p\}$

```
S ← 1 − M_bs
segment I into two regions I_1 and I_2 by Ncut algorithm [135]
O ← {I_1, I_2}
while O ≠ ∅:
    for each R ∈ O:
        select training set D'_R according to M_bs
        train FE-SAE
        for each superpixel p ∈ R:
            calculate saliency value s'_p by (3.20)
            s_p ← (s_p + s'_p)/2
        end for
        remove R from O
        if 0.3 × |R| ≤ |D'_R| ≤ 0.7 × |R| then:
            segment R into two regions R_1 and R_2 by Ncut algorithm
            O ← O ∪ {R_1, R_2}
        end if
    end for
end while
```

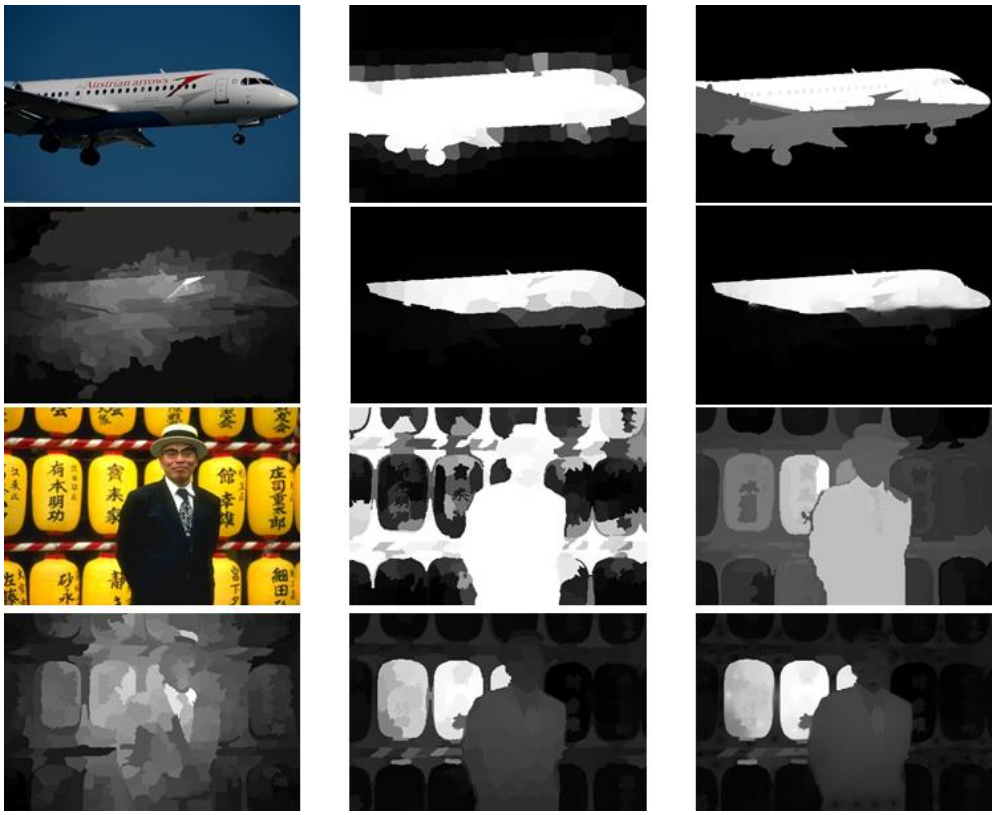
3.3.2. Experimental results

For BS-SAE model, we stack three AEs to extract feature representation in high-level manners, with 7000, 3500 and 2000 hidden nodes in each AE, respectively. As the MSRA-10K [21] dataset provides a large variety of natural images and the corresponding pixel-wise saliency annotations, we randomly selected 9000 images from the dataset to train the BS-SAE and left out 1000 images for use in the validation. As suggested in [20, 126], before input to BS-SAE, $f(p_{bs})$ is corrupted to enhance the robustness across a large training set, in which some of the units are set to be zero randomly. For FE-SAE model, we stacked two AEs to boost the performance of data reconstruction, with 60 hidden nodes in each of the AE. As the number of training samples is small (generally less than 250), we did not corrupt the original input vector in FE-SAE to make the trained model more specific to the small training set. The two models were both implemented with Theano frame [136, 137], which enabled the use of GPU to boost the speed in the training phase. The hyperparameters in the training of BS-SAE and FE-SAE are listed in Table 3.3.

Table 3.3 The hyperparameters in the training of two models.

	BS-SAE		FE-SAE	
	Pre-training	Fine-tuning	Pre-training	Fine-tuning
Training epoch	15	60	15	100
Learning rate	1e-2	1e-6 in first 20 epochs; 8e-8 in last 40 epochs.	1e-2	1e-3

Figure 3.8 visually depicts that our proposed background search and foreground estimation method (BSFE) achieves best qualitative performance against comparison methods. For example, as shown in the first row, BSFE successfully recognized the whole saliency object while most of the other methods only recognized the main body of the airplane but failed to capture the wing and the landing gears. Such favorable performance is largely attributed to the BS-SDAE, as it can semantically infer the whole structure of the airplane from the learned features. Similarly in the fifth row, contrary to our method which accurately recognized the bicycle and the child as the salient objects, even the boundary-background priors based comparison methods (e.g. LR12 and MC13) failed to capture the bicycle which covers and in contact with the bottom of the image.



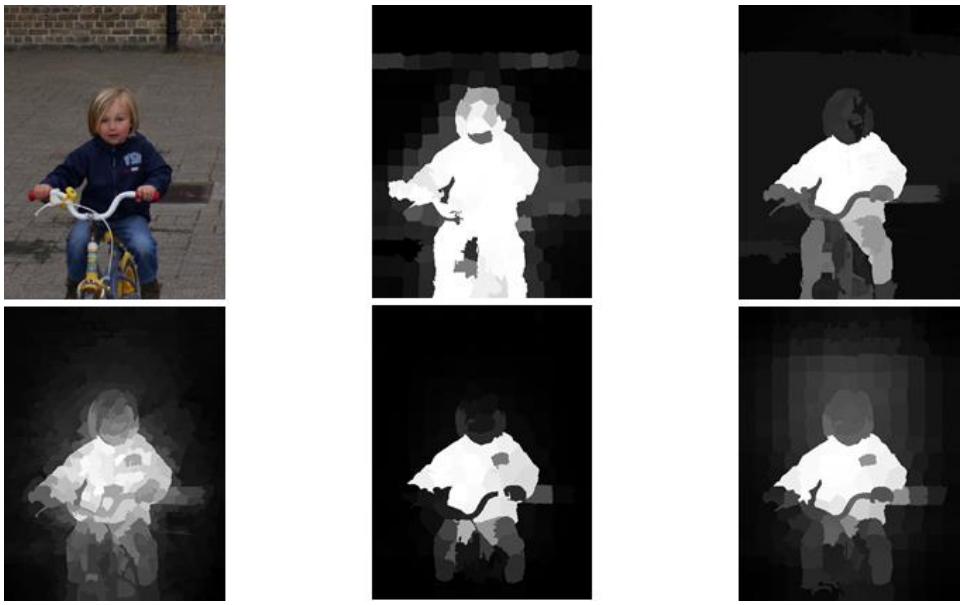


Figure 3.8 Example results of our proposed BSFE method. From left to right in the first, third and fifth row: original images, saliency maps by BSFE and HS13 [22]. From left to right in the second, forth, and sixth row: saliency maps by LR12 [138], MC13 [17] and RR15 [13].

3.3.3. Evaluation

We evaluated our proposed algorithm on four public benchmark datasets, i.e. ECSSD [22], PASCAL-S [23], SED1 [139] and SED2 [139]. Six popular state-of-the-art algorithms which employ image boundaries as background seeds were chosen as comparison methods, including RR15 [13], HS13 [22], MC13 [17], MR13 [14], FT09 [140] and LR12 [138]. Following [8, 13, 15], we adopt F-measure (FM), precision-recall (PR) curve and mean absolute error (MAE) [13] to evaluate the performances. The evaluation results shown in Figure 3.9 quantitatively demonstrate the superiority of our method on most datasets. Note that our BSFE method even achieved double-best results in terms of FM and MAE on PASCAL-S and SED2 datasets which contain more challenging scenarios with complex structures and double-salient-objects.

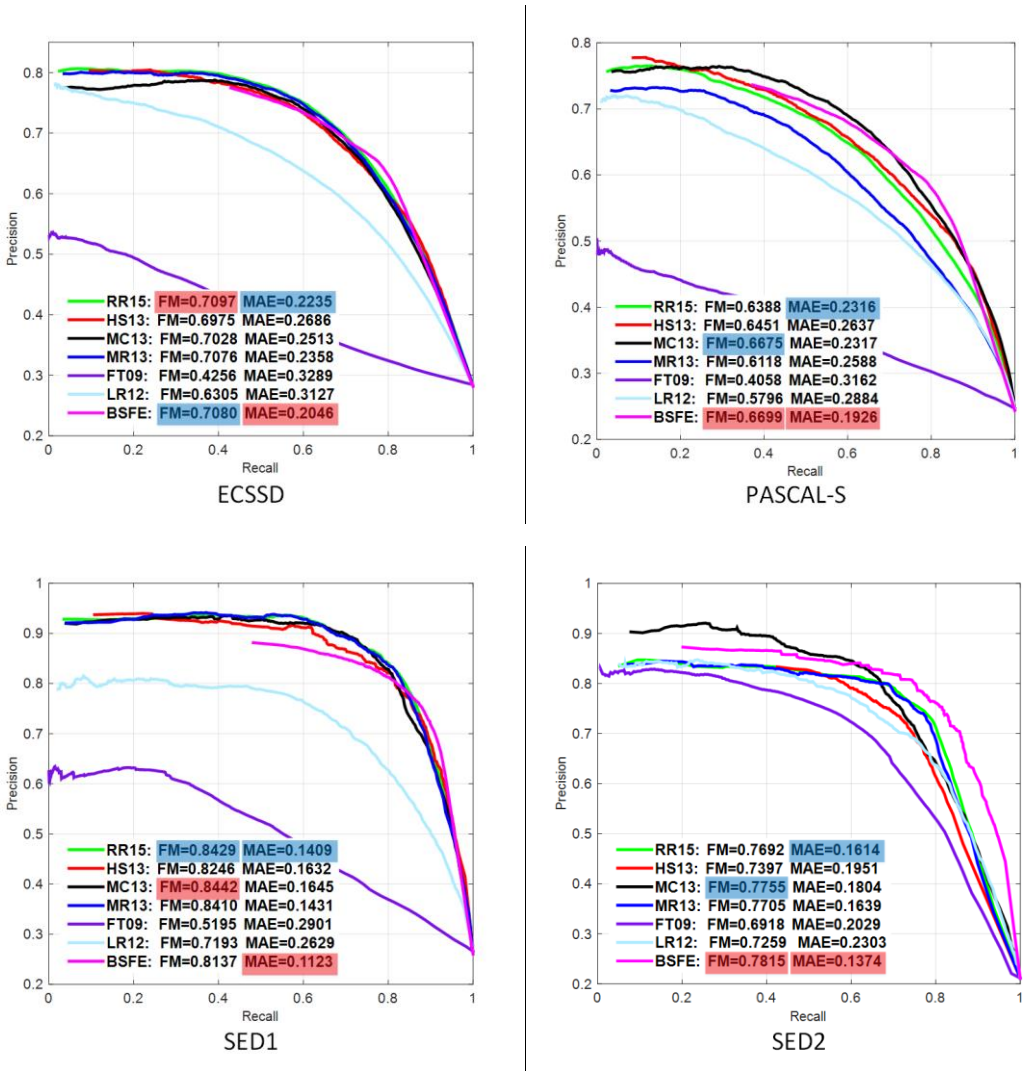


Figure 3.9 The PR curve, FM and MAE of benchmarking methods on four public datasets. The best and second best results are padded with red and blue rectangle respectively.

As convolutional neural networks (CNN) is powerful for feature extraction and data analysis, we also compared our proposed algorithm with a CNN based method (MDCL [45] proposed in 2015). The qualitative comparison and quantitative comparison in terms of F-measure and MAE are shown in Figure 3.10, Table 3.4 and Table 3.5 respectively. The comparison results shows that our BSFE method cannot achieve the better performance of CNN based method. The main reason is that compared to CNN, SAE is not sufficient to extract high-level features with relative shallow layers and may loss original spatial information during input vectorization.

However, as shown in Figure 3.7 and Figure 3.9, our BSFE method can still provide the more meaningful and reliable prior seeds and thus boost the final recognition results according to the comparisons with conventional boundary-seeds methods.

Table 3.4 F-measure of our BSFE method and CNN based method (MCDL) on benchmarking datasets.

	PASCAL-S	ECSSD	SED1	SED2
BSFE	0.6699	0.7080	0.8137	0.7815
MDCL	0.6998	0.7469	0.8581	0.7847

Table 3.5 MAE of our BSFE method and CNN based method (MCDL) on benchmarking datasets.

	PASCAL-S	ECSSD	SED1	SED2
BSFE	0.1926	0.2046	0.1132	0.1374
MDCL	0.1597	0.1752	0.0875	0.1074

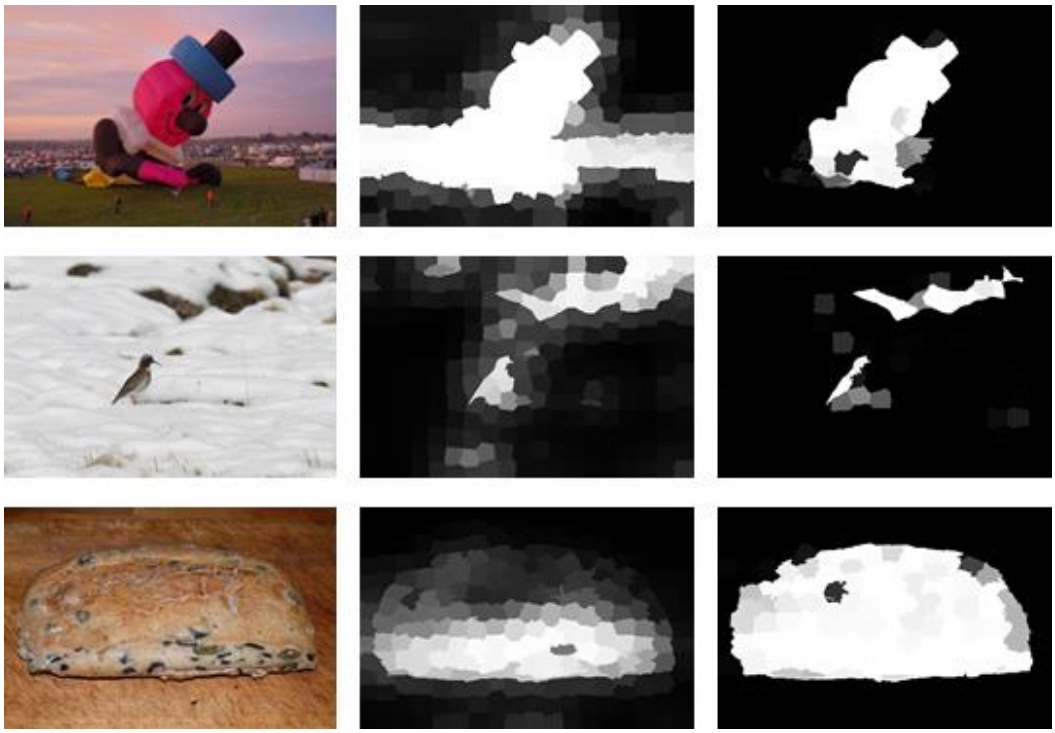


Figure 3.10 Example results of our BSFE method and CNN based method (MCDL). From left to right: original images, the saliency maps produced by BSFE and MCDL.

3.4. Summary

In this chapter, we have proposed an automatic prostate recognition method on MR images based on SAE. Compared to the most existing works with AE, we let the SAE itself serve as a classifier to focus on the prostate feature extraction. An image energy minimization scheme is then proposed to optimize the prostate recognition map constructed by SAE. Our method is compared against three benchmark classifiers and atlas-based seeds-selection approach on the PROMISE12 database, demonstrating superiority in both PR curves and F-measures. Furthermore, we have also extended the AE-based prostate recognition model for the aim of saliency object detection, and achieved competitive results on popular public datasets.

4. A general object recognition model via multi-contexts combined deep neural networks

Object recognition has been extensively studied in many works. Specifically, for saliency object detection, most conventional methods form a rough saliency map with various prior knowledge, such as flash cues [141], boundary-background priors (image boundaries are treated as background seeds) [13-15, 17, 20, 133, 134], and dark channel [15], and then construct the final saliency map. However, these priors are not always reliable. For example, in the first row of Figure 1.3, the plants at the bottom side of the image 'pops' out, compared to the consistent regions at the other sides of the image, and thus tends to be labeled as false positive areas. To lessen such negative impacts, Na et.al. [15] proposed a novel method which estimates saliency value with supported vector machine directly learned from the tested image itself. Such self-shallow-learning approaches significantly improve the performance compared to those none-learning approaches [17, 133]. However, the methods are often limited by the insufficient number of training samples. It may fail to construct a strong classifier for the challenging images. Different from the aforementioned approaches, the works of [13, 14, 142] adopt a certain number of seeds over the image to infer the remaining unlabeled pixels by formulating energy minimization scheme, which less rely on the prior knowledge. However, as these algorithms [13, 14, 142] directly use low level-cues, e.g. RGB/CIELab color values, to estimate the distance among pixels, such labelling propagation approach may produce stochastic result for the image where the foreground and background share similar appearance.

As for anatomical structure recognition on biomedical images, many works highly on atlas maps to register a specific image. However, the errors produced by atlas maps still impede precise recognition of anatomical structure from biomedical images.

To address the aforementioned issues, we proposed a general deep networks based model for object recognition on natural and biomedical images in multi-scales. Different from other multi-scales methods [45], we specially designed the structure of input data so that the model can infer the relations among different scales automatically.

4.1. Object recognition method

Our proposed multi-contexts combined convolutional neural networks (MCDN) for object recognition is superpixel-wised. Superpixel algorithms, such as NC [143], FH [144], QS [145] and SLIC [123], tend to cluster pixels perceptually which serves as the atomic regions in many computer vision tasks. Following the most saliency detection works [14, 15, 38, 134], we adopt SLIC algorithm to partition the image into N_{sp} non-overlapping superpixels.

4.1.1. Input data preparation

The existing works [8, 45] integrate independent DNNs with handcrafted functions to exploit saliency map in multi-scales. However, the relatively small number of parameters in handcrafted functions sometimes cannot well depict the correlations of multi-scaled intermediates, thus may cognitively and semantically conflict with the intentions of DNNs. Comparably, we construct a uniform DNN to learn such integration automatically and adaptively, with the concatenation of the multi-contexts. We first construct the input data to feed DNN.

For a superpixel p of an image I with three channels, we extract the corresponding local-context and global-context. The local context of p is the region composed by itself and its neighboring superpixels. The global context of p is the image I. Thus, the input data F_p of DNN, with respective to the superpixel p, is composed of local-context F_p^{local} and global-context F_p^{global}:

$$F_p = \begin{bmatrix} F_p^{local} \\ F_p^{global} \end{bmatrix} \tag{4.1}$$

In order to mark p on I, the values at the region of p are set to zero on the global context. As the global context can infer the region of p from the local context, there is no color information loss during such padding operation.

4.1.2. Convolutional neural networks training

Due to the broadly-validated steady performance, we adopt the AlexNet model [42] to construct DNN in this work, with a softmax regression at the top of the last network layer to estimate the probabilities of suerpixels being salient. In order to achieve non-linear transformation, the rectified linear unit (ReLU) [146] is utilized in the proposed DNN structure. Additionally, the batch normalization [147] is inserted following each convolutional layer (except the last layer), which boosts the training phase with high learning rate and lessens the impacts by weights initialization. Table 4.1 is the detailed structure of this deep networks.

Table 4.1 The detailed structure of our proposed deep network. c: convolutional layer; b: batch normalization layer; r: ReLU layer; f: fully connected layer; s: softmax regression layer.

Layer	Type	Channel	Filter size	Conv. stride	Conv. pad	Pooling size	Pooling stride	Input size
1	c+b+r	96	11×11	4	0	3×3	2	227×227×6
2	c+b+r	256	5×5	1	2	3×3	2	55×55×96
3	c+b+r	384	3×3	1	1	N/A	N/A	27×27×256
4	c+b+r	384	3×3	1	1	N/A	N/A	13×13×384
5	c+b+r	256	3×3	1	1	3×3	2	13×13×384
6	c+b+r	4096	6×6	1	0	N/A	N/A	6×6×256
7	f+b+r	4096	1×1	1	0	N/A	N/A	1×1×4096
8	f+s	2	1×1	1	0	N/A	N/A	1×1×4096

Given the training superpixel set $\mathcal{A} = \{p\}$ and the corresponding label set $\mathcal{B} = \{b_p\}$, the input data set $\{F_p\}$ can be obtained as described in chapter 4.1.1. The local-context and global-context in each F_p are both resized to 227×227×3 to fit the structure of proposed deep network, thus we have $F_p \in \mathbb{R}^{227\times227\times6}$. The deep network is then trained by the training set $\mathcal{A}$, with the aim of minimizing the following cost function:

$$J(\theta) = -\frac{1}{|\mathcal{A}|}\sum_{p\in\mathcal{A}}\sum_{j=0}^{1} T\left(b_p = j\right)\log P\left(b_p = j\middle|x\right) + \frac{\lambda}{2}\sum_{z=1}^{Z}\theta_z^2 \tag{4.2}$$

$$P\left(b_p = j\middle|x\right) = \frac{\exp(x_j)}{\sum_{i=0}^{1}\exp(x_i)} \tag{4.3}$$

where $T(\cdot)$ is 1 if the condition inside the parentheses is true and 0 otherwise; x is the output of the penultimate layer; $P\left(b_p = j\middle|x\right)$ is the probability labeling b_p as j; λ

is a fixed parameter and set to 0.0005 empirically; Z is the total number of layers in the network; and θ_z is the weight of the z-th layer. The second term of (4.2) is to balance the first term such that it restricts θ from growing too large unless necessary [148], thus improves the generalization of the trained network. To enable backpropagation of $J(\theta)$, we calculate the partial derivative

$$\frac{\partial}{\partial x_j} J(\theta) = -\frac{1}{|\mathcal{A}|} \sum_{p \in \mathcal{A}} (T(b_p = j) - P(b_p = j|x)) \tag{4.4}$$

which allows the loss gradient to flow back to the former layers and thus updates θ by the gradient descent optimization algorithm [127] in iterations.

4.1.3. Superpixel classification

In testing phase, the image is first partitioned into N_{sp} superpixels and then extracted corresponding input data by (4.1) for each superpixels. With the well-trained multi-contexts combined DNN, we can predict the likelihood of each superpixel belonging to the class j by (4.3), where the value 1 of j is foreground and 0 is background. Afterwards, $P\left(b_p = 1\middle|x\right)$ can be utilized to estimate the saliency value of the superpixel.

4.2. Validation of saliency object detection

4.2.1. Setup and dataset

As the MSRA-10K [21] dataset covers a large variety of scenarios with pixel-level saliency annotations, in the experiment on saliency object detection, we randomly select 9,000 images from it to compose the training set to train the proposed multi-contexts combined DNN, and leave 1,000 images for validation. The aim of validation is to evaluate the performance of the current trained DNN

following each training epoch, and do not update the learnable parameters in DNN. The algorithm was implemented with MatConvNet framework [149] and the training process for the DNN was conducted on a PC with Intel 6-Core i7-5820K 3.3GHz CPU, 64GB RAM and a GeForce GTX TITAN X 12GB GPU. Other detailed hyperparameters in the training phase of the DNN are listed in Table 4.2.

Table 4.2 The hyperparameters in the training phase of the DNN for saliency object detection.

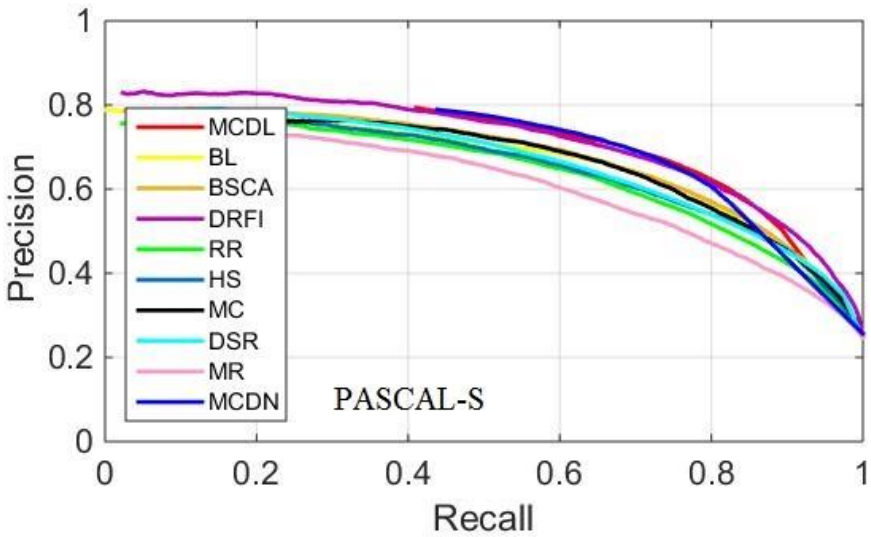
N_{sp}	Batch size	Momentum	Training epoch	Learning rate
200	200	0.9000	20	20-point logarithm space between 0.1 to 0.0001

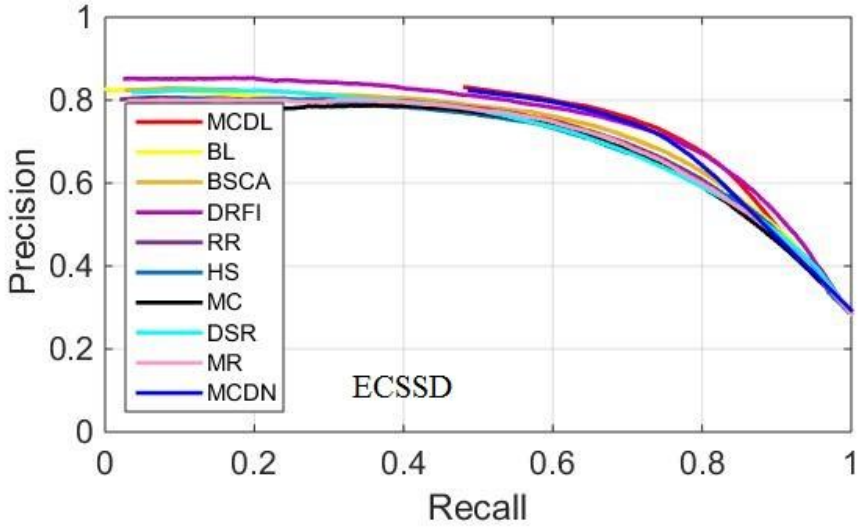
In testing phase, we run our proposed algorithm on five benchmark datasets, i.e. PASCAL-S [23], ECSSD [22], SED1 [139], SED2 [139] and DUT-OMRON [14]. PASCAL-S contains 850 natural images which are built for the validation of the PASCAL VOC 2010 segmentation challenge with complex structures. ECSSD contains 1,000 images from the Internet. SED1 contains 100 single-salient-object images, while SED2 contains 100 double-salient-object images that is more challenging compared to SED1. DUT-OMRON contains 5,168 images with more challenging scenarios compared to the aforementioned four datasets. The pixel-wise ground truth masks of all the images on the five datasets are available by manual segmentations.

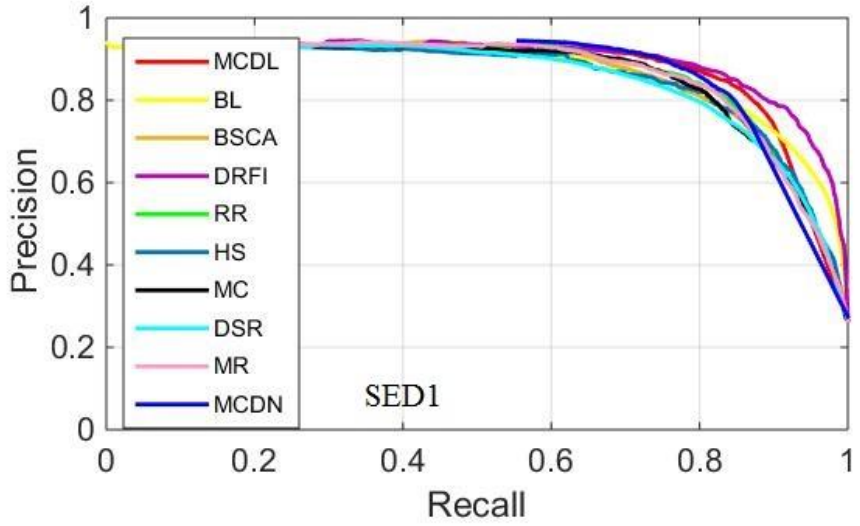
4.2.2. Experimental results

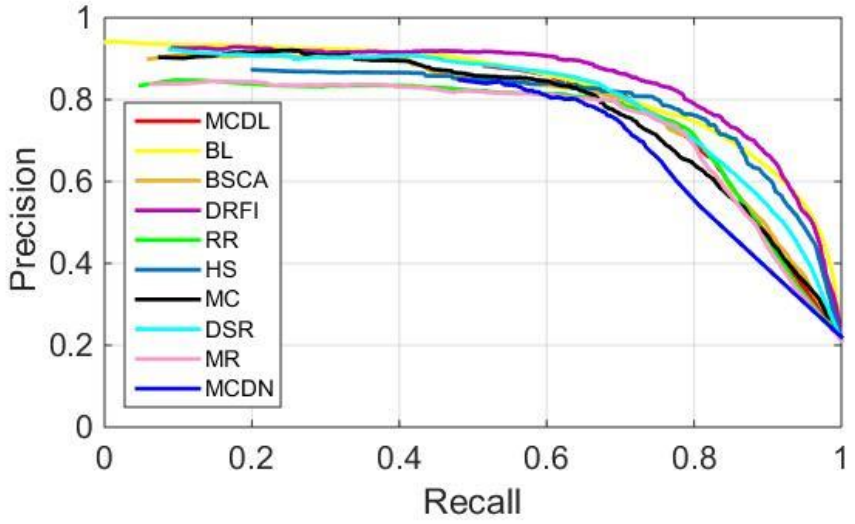
We evaluate our proposed MCDN method against nine state-of-the-art methods, including MCDL [45], DRFI [150], BL [15], MC [17], MR [14], RR [13], HS [22], BSCA [12] and DSR [151] on PASCAL-S [23], ECSSD [22], SED1 [139], SED2 [139] and DUT-OMRON [14] datasets respectively. The comparison methods are set by default parameters published in their original papers or codes, and are conducted under the same environment. The experimental results, in terms of PR curve, F-measure and MAE, are quantitatively shown in Figure 4.1, Table 4.3 and

Table 4.4 respectively.




Precision
Recall
MCDL
BL
BSCA
DRFI
RR
HS
MC
DSR
MR
MCDN
ECSSD


Precision
Recall
MCDL
BL
BSCA
DRFI
RR
HS
MC
DSR
MR
MCDN
SED1


Precision
Recall
MCDL
BL
BSCA
DRFI
RR
HS
MC
DSR
MR
MCDN

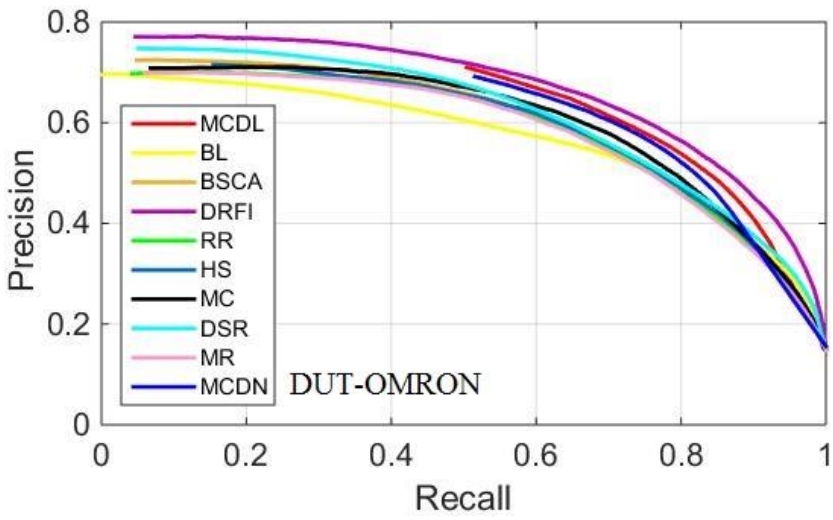


Figure 4.1 PR curves of our method (MCDN) and the counterparts.

Table 4.3 F-measure of our method (MCDN) and the counterparts. The best and second best results are shown in red and blue.

	PASCAL-S	ECSSD	SED1	SED2	DUT-OMRON
MCDN	0.7041	0.7430	0.8619	0.7575	0.6444
MCDL	0.6998	0.7469	0.8581	0.7847	0.6509
BL	0.6228	0.7161	0.8404	0.7934	0.5798
BSCA	0.6694	0.7180	0.8319	0.7797	0.6171
DRFI	0.6938	0.7358	0.8638	0.8226	0.6640
RR	0.6388	0.7097	0.8429	0.7692	0.6127
HS	0.6451	0.6975	0.8246	0.7815	0.6161
MC	0.6675	0.7028	0.8442	0.7755	0.6273
DSR	0.6506	0.6986	0.8186	0.7868	0.6269
MR	0.6188	0.7076	0.8410	0.7705	0.6108

Table 4.4 MAE of our method (MCDN) and the counterparts. The best and second best results are shown in red and blue.

	PASCAL-S	ECSSD	SED1	SED2	DUT-OMRON
MCDN	0.1625	0.1813	0.0911	0.1279	0.1166
MCDL	0.1597	0.1752	0.0875	0.1074	0.1183
BL	0.2493	0.2620	0.1900	0.1403	0.2401
BSCA	0.2238	0.2235	0.1548	0.1583	0.1908
DRFI	0.2098	0.2256	0.1485	0.1403	0.1496
RR	0.2316	0.2235	0.1409	0.1614	0.1845
HS	0.2637	0.2686	0.1632	0.1951	0.2274
MC	0.2317	0.2513	0.1645	0.1804	0.1863
DSR	0.2079	0.2263	0.1599	0.1894	0.1388
MR	0.2588	0.2358	0.1431	0.1639	0.1868

According to Table 4.3, although DFRI performs best on three datasets, our proposed MCDN method ranks top-2 on four out of five datasets, which proves the robustness of MCDN. The two deep networks based methods (i.e. MCDN and MCDL) place the top-2 positions on all five datasets, in terms of MAE. However, MCDL beats our method on four datasets. We will discuss this competition and the results in Chapter 4.2.3. From the examples in Figure 4.2, not only does MCDN recognize the rough positions and shapes of the salient object, but also can well suppress the background.

Figure 4.2 Saliency example maps by our method and conventional methods. From top to bottom: original images, saliency maps produced by our method (MCDN), BL [15] and MR [14].

4.2.3. Comparison with other deep networks based method

In this chapter, we evaluate the performances of MCDN and other deep networks based saliency object detection methods. Typically, we choose one of the state-of-the-art counterparts, i.e. MDCL [45] proposed in 2015, as the comparison. We conduct our method and MCDL with the same environment, including training set, training epoch and learning rate. Both are implemented with AlexNet model.

As shown in Figure 4.3, compared to the results by MCDL, MCDN can produce more smoothed saliency maps. Our higher performance than MCDL is attributed to two aspects. Firstly, MCDL puts the to-be-classified superpixel at the center of the image but does not precisely mark it. In contrast, our method directly marks the to-be-classified superpixel for DNN. As our DNN can precisely locate the to-be-classified superpixel, it outperforms MCDL. Secondly, as discussed in Chapter 4.1.1, the combination of mulit-context by deep learning also attributes to the better performance of our method. Noted that MCDL beats our method in terms of MAE. The reason is that as our method intends to produce more smoothed inner regions and suppressed background, the false-positive and false-negative results greatly increase MAE score of our method.

Figure 4.3 Saliency example maps by our method and deep networks based methods. From top to bottom: original images, saliency maps produced by our method (MCDN) and MCDL [45].

4.3. Validation of prostate recognition

4.3.1. Setup and dataset

We used prostate MR Image Segmentation 2012 (PROMISE12) dataset [129]. This set contains 50 cases which are from multi-center and multi-vendor, and with different acquisition protocols [89]. Each case comprises of a set of transversal T2-weighted MR images, and pixel-wised prostate annotations by experts.

Different from the natural images, the prostate MR image is one-channel-intensity image so that the input data F_p is two dimensions. The other settings for saliency object detection (Chapter 4.2), including the hyperparameters of DNN and PC configurations, are shared here.

In the application of prostate recognition, we applied ten-fold cross validation on the dataset. Specifically, the dataset was randomly partitioned into ten groups; nine groups constituted the training set and the left one group was used for testing. We performed such validation in iterations until all the groups were tested.

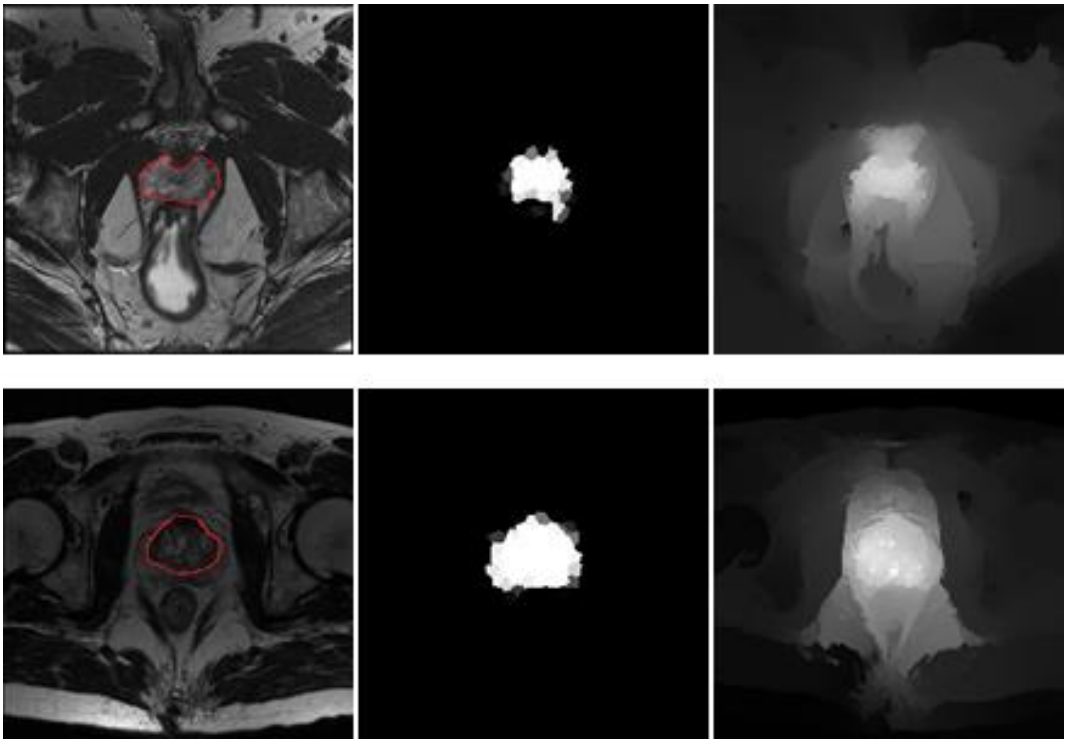
4.3.2. Experimental results

Atlas probability maps provide favorable foreground priors and are widely applied in many prostate segmentation algorithms [49, 50] for seeds selection. We compare our MCDN method with the atlas-based seeds-selection proposed in one of the prostate segmentation work (RW) [49]. We use the default parameter settings in [49] for comparison. The MCDN method strongly beats the atlas-based seeds-selection, both in terms of precision and F-measure. Figure 4.4 visually shows the significant superiority of our method. Although the atlas-based seeds-selection in [49]

can recognize the prostate, it leads to more false-positive cases compared to our method.

Table 4.5 Precision and F-measure of our method (MCDN) and atlas-based seeds-selection (RW).

	MCDN	RW
Precision	0.9285	0.8284
F-measure	0.8383	0.6617



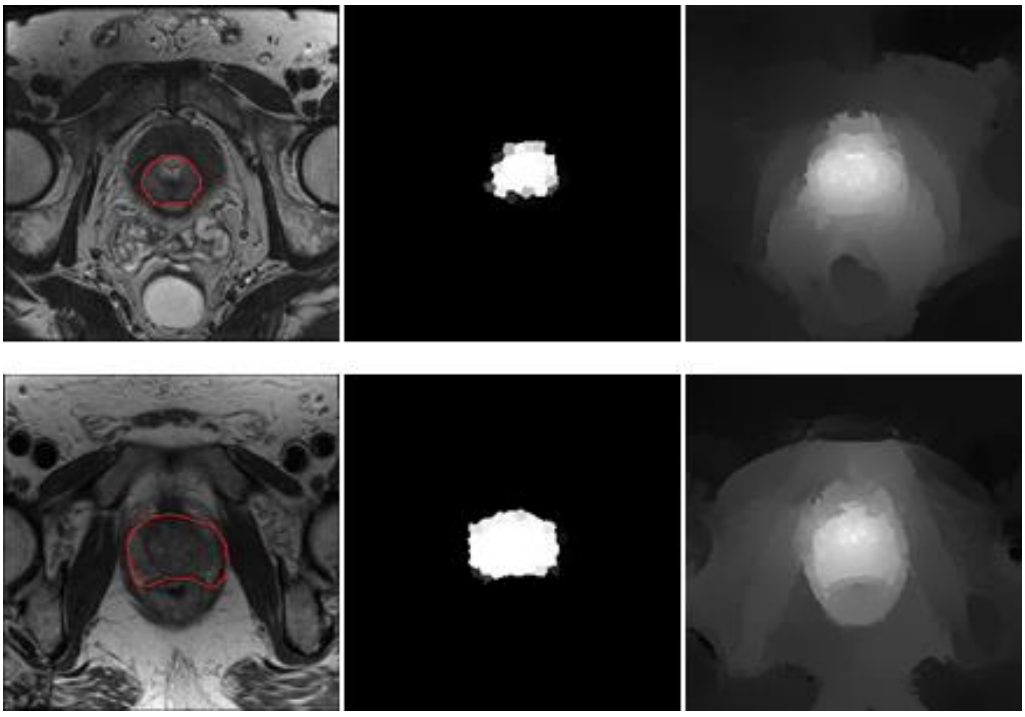


Figure 4.4 Comparisons of our method (MCDN) and atlas-based seeds-selection in prostate recognition. From left to right: original prostate MR image (prostate regions are delineated in red contours), recognition results by our method and RW [49].

4.4. Summary

In this chapter, we have proposed a general deep neural networks based method for object recognition on natural and biomedical images. By integrating the local and global contexts in input data, our model extracts high-level features in multi-scales thus achieves better results compared to conventional methods with handcrafted features. Experimental validation on saliency object detection and prostate recognition demonstrated that our model is robust to different types of object recognitions across various datasets.

5. A novel saliency image energy cooperating with region priors

Many tasks such as image segmentation [152, 153], restoration [154], object recognition [155], and texture synthesis [156] can be solved through the optimization of image energy functions constructed in variety of ways. It can be effectively and efficiently minimized by max-flow algorithm [95] to solve binary labeling task. Saliency detection with energy minimization [13, 14, 142] has been studied for many years. Wei et al. [142] define geodesic saliency to form image energy and apply Dijkstra's algorithm to discover the shortest path over the image from background to foreground. The works of [13, 14] are based on manifold ranking, which minimize the defined energy by differential method. In order to obtain more precise saliency maps, in this chapter, we propose a novel saliency image energy and the refined saliency maps can be formed by minimizing the proposed image energy. The region priors are imposed on the image energy to guide the recognition and segmentation of saliency objects on the basis of our proposed three observations.

5.1. Method

Our proposed saliency image energy is composed of smooth penalty and data penalty. The aim of smooth penalty is to encourage smooth inner regions and distinct region boundaries. Instead of exploiting the distance of pixels with color appearance, we adopt image segmentation approach to generate pre-segments and use them as region priors over the image, and then formulate smooth penalty on that basis. The data penalty represents the label-preferences of pixels, which can be directly estimated by the saliency map from most conventional approaches, e.g. [14, 17].

However, as described in Chapter 4, the conventional approaches may fail to assign precise labels to pixels on complex images, thus eventually deteriorate the performance of the whole image energy in saliency detection. To achieve better performance, we adopt the saliency map generated by the proposed multi-contexts combined DNN for more reliable label-preferences as the data penalty. The labels for saliency detection can be thus assigned to each pixel by finding the minimum solution for the image energy. In this way, the deep networks based image energy (DNIE) is our proposed new saliency object detection method. The pipelines of the DNIE approach for saliency object detection is shown in Figure 5.1. In the remaining parts of this chapter, we mainly focus on the formulation of the smooth penalty and the method to produce the final saliency map according to the proposed saliency image energy.

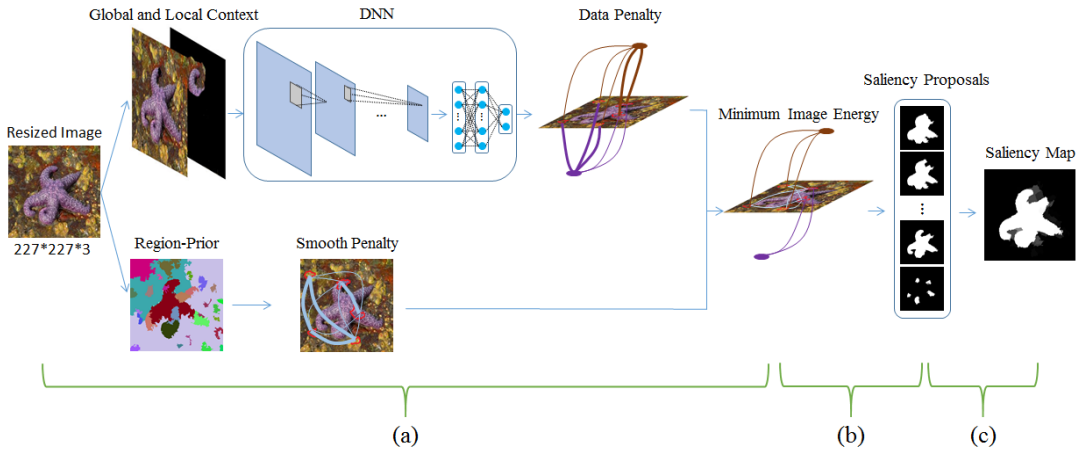


Figure 5.1 The pipelines of our proposed DNIE approach for saliency detection. (a) Image energy construction with data penalty by multi-contexts combined DNN model (Chapter 4) and smooth penalty by region-priors. (b) Image energy minimization for saliency proposals. (c) Saliency estimation. In DNN model, the third dimensions of layers are visually omitted in this figure.

5.1.1. Problem formulation

We use superpixel by SLIC algorithm [123] as the basic homogenous region in the further operations. Formally, an image I is partitioned into a superpixel set $\mathcal{P} = \{p_1, p_2, \dots, p_N\}$ with N elements, where we always ignore the image notation I for simplification. The saliency detection on image I aims to find a labeling configuration $\mathcal{L} = \{l_{p_1}, l_{p_2}, \dots, l_{p_N}\}$ for each superpixel p_i in $\mathcal{P}$, where $l_{p_i} \in \{0,1\}$ represents background and foreground respectively. $\mathcal{L}$ is then transferred into a soft labeling configuration $\mathcal{L}^* = \{l^*_{p_1}, l^*_{p_2}, \dots, l^*_{p_N}\}$ to estimate the probability of p_i being salient. For brevity, we omit the subscript to notate the element in a set such that p is the general notation of p_i in $\mathcal{P}$.

A proper labeling configuration should appropriately maintain the individual label-preferences of superpixels by observation or pre-specified likelihood function, and meanwhile tend to produce the smooth saliency region and suppressed background. Based on this motivation, we find the best labeling configuration by minimizing the following image energy:

$$E(\mathcal{L}) = (1 - w) \sum_{p \in \mathcal{P}} D(l, p) + w \sum_{p,q \in \mathcal{P}} V\left(l_p, l_q, p, q\right) \qquad (5.1)$$
$$s.t.\ l_p + l_q = 1$$

where $D(l, p)$ is the data penalty to assign label l to p based on the label-preferences and $V\left(l_p, l_q, p, q\right)$ is the pairwise smooth penalty to assign different labels l_p, l_q to p, q respectively. The weighting factor w controls the weight between these two terms in (5.1).

5.1.2. Data penalty

The data penalty initializes the label-preferences of a superpixel, which makes the image energy a task-driven (i.e. saliency-driven) scheme. In DNIE approach, the saliency map generated by multi-contexts combined DNN is adopted to estimate the data penalty:

$$D(l,p) = P_p(l = 1|\theta) \cdot T(l = 1) + (1 - P_p(l = 1|\theta)) \cdot T(l = 0) \tag{5.2}$$

where $T(\cdot)$ is 1 if the condition inside the parentheses is true and 0 otherwise; $P_p(l = 1|\theta)$ has been defined in chapter 4.1.

In addition to the multi-contexts combined DNN, our proposed saliency image energy can also adopt other types of saliency maps produced by conventional low-cues based methods, such as MR [14] and MC [17]. The data penalty of p corresponding to the saliency map Sal is

$$D(l,p) = Sal(p) \cdot T(l = 1) + (1 - Sal(p)) \cdot T(l = 0) \tag{5.3}$$

5.1.3. Smooth penalty

The smooth penalty estimates the cost of assigning pairwise superpixels with different labels. In DNIE approach, since the data penalty is produced by the high-level image representations from DNN, which is superior to those low-level cues [8, 45], the smooth penalty aims to generate results in accordance with the data penalty. It tends to separate saliency objects from background with clear region appearance, integrating with the data penalty. Instead of merely using low-level cues, we adopt image segmentation algorithm to generate region-prior to explore the differences among superpixels in calculating smooth penalty as follows. This is different from

some traditional region-smoothness works [15, 157], which simply rely on low-level cues.

The image I is segmented into M regions, denoted as $\mathcal{R} = \{r_1, r_2, \dots, r_M\}$, via graph-based segmentation algorithm [144], and the region-prior $\mathcal{O} = \{\sigma_p \in \mathcal{R}\}$ of I is then generated, where σ_p is the region containing superpixel p. With the region-prior $\mathcal{O}$, for saliency detection, the smooth penalty should follow the three observations:

Observation 1. For common images, the probability of a pair of superpixels sharing a same label is increasing, with their position distance decreasing.

Observation 2. A pair of superpixels belonging to different regions, with a large position distance, often tend to take different labels, as the saliency region is compact in most cases.

Observation 3. In the same region, especially the simple and consistent region (e.g. sky and ocean), the superpixels with similar appearance often share same labels.

The observation 1 and 2 are fundamental guides to the definition of smooth penalty, and the observation 3 is an additional one improving the labeling results over the whole image energy. Therefore, based on these observations, for a pair of superpixels p, q in image I, the smooth penalty $V(l_p, l_q, p, q)$ is defined as

$$V(l_p, l_q, p, q) = \frac{T(\sigma_p \neq \sigma_q)}{1 + G(p,q) + G(\sigma_p, \sigma_q)} + \frac{T(\sigma_p = \sigma_q)}{1 + \left\|c_p - c_q\right\| \cdot T(H(p) = H(q)) + G(p,q) \cdot T(H(p) \neq H(q))} \tag{5.4}$$

$$s.t.\ \ l_p + l_q = 1$$

We explain (5.4) in details as follows. The subjection term in (5.4) ensures that the smooth penalty is the cost of assigning different labels to pairwise

superpixels. The first term of (5.4) is the external-region contrasted penalty which estimates the cost of assigning labels to pairwise superpixels belonging to different regions, while the second term of (5.4) is the inner-region smooth penalty which determines the cost of labeling pairwise superpixels within same region. The position distance term $G(u, v)$ captures the Observation 1, as a small $G(u, v)$ leads a large smooth penalty encouraging same label to pairwise superpixels with small position distance. In accordance with Observation 2, the region-difference term $G(\sigma_p, \sigma_q)$ is imposed on the first term of (5.4) to discourage the assignment of same labels to superpixels within different regions. Inspired by Observation 3, the calculation of the RGB distance for superpixels with similar appearance is added to the second term of (5.4) to improve the results with smooth saliency region and suppressed background.

5.1.4. Saliency proposals and estimation

Given the exposition of data penalty and smooth penalty, the image energy can be formulated by (5.1) and the labeling configuration $\mathcal{L}^{(w)}$ can then be determined by minimizing (5.1) under a specific weighting factor w. In this work, we adopt the max-flow algorithm in [95] to minimize (5.1). To achieve better performance, instead of using fixed weighting factors in (5.1) to control the weights among data penalty and smooth penalty [157], we generate saliency proposals $\{\mathcal{L}^{(w)}\}$ under different values of w and then integrate $\{\mathcal{L}^{(w)}\}$ to $\mathcal{L}$ as the final labeling configuration of the image energy minimization scheme, with the normalized coefficients by Gaussian function.

As the labeling configuration $\mathcal{L}$ is a binary result, similar to [15], we linearly combine $\mathcal{L}$ with $\{P_p(l = 1|\theta)\}$ that is defined in chapter 4.1 to obtain a soft labeling configuration $\mathcal{L}^*$ as follows

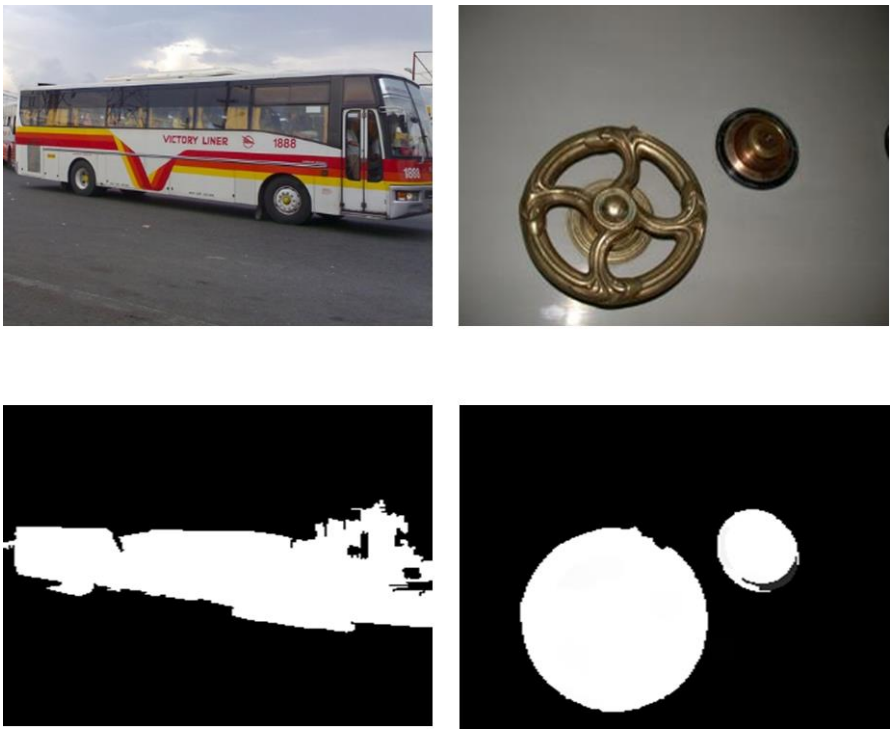
$$L^* = \{l_p^*\} = \{\frac{l_p + P_p(l = 1|\theta)}{2}\} \tag{5.5}$$

where $l_p \in \mathcal{L}$ corresponds to the superpixel p in $\mathcal{P}$ of the image I. Then we estimate the probability of superpixel p being foreground as l_p^*, and the saliency map can thus be obtained.

5.2. Experiment and evaluation

5.2.1. Overall performance of DNIE

Some example saliency maps by DNIE are shown in Figure 5.2 in which DNIE has a better visually results. For example, while low-cues based approach cannot recognize the bus on the image of the first column of Figure 5.2, DNIE enables to capture the parts of the bus. Although DNN based approach can also recognize the same parts of the bus, the inner regions are not as smooth as the recognized regions by DNIE. This superiority is attributed to the proposed smooth penalty in the saliency image energy.



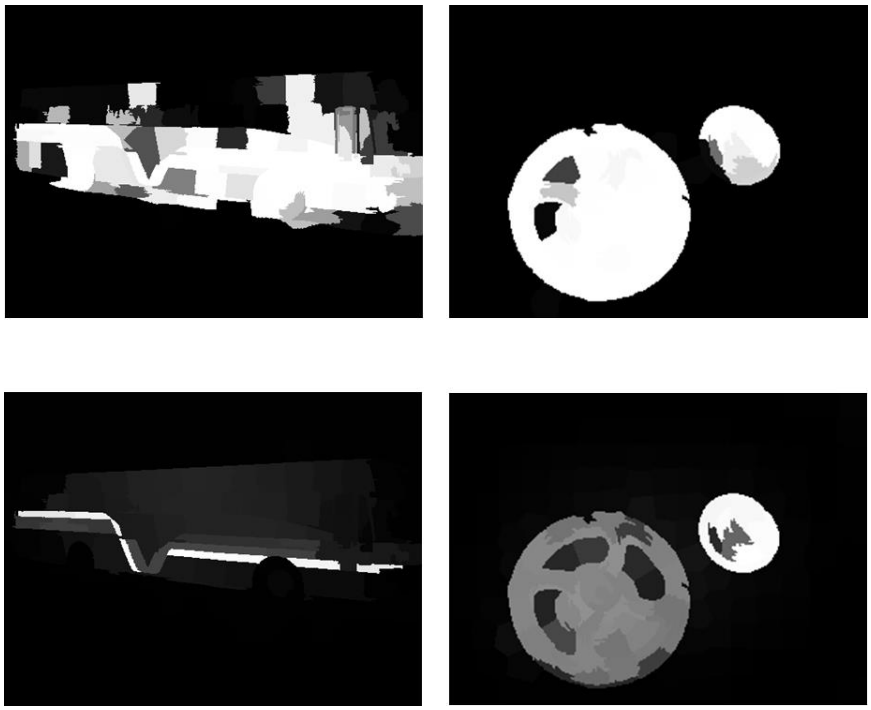
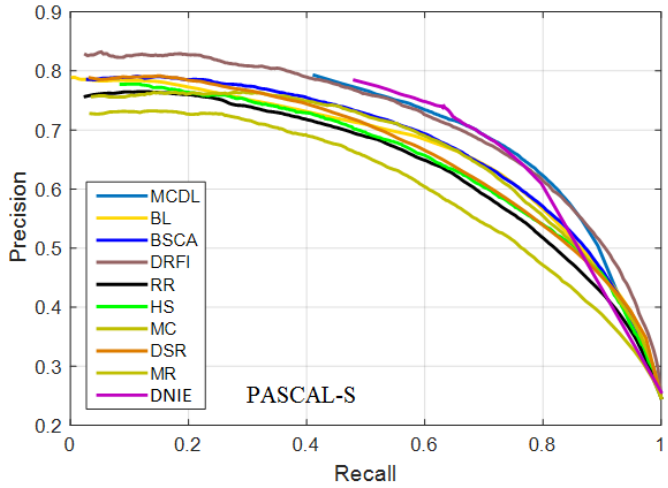
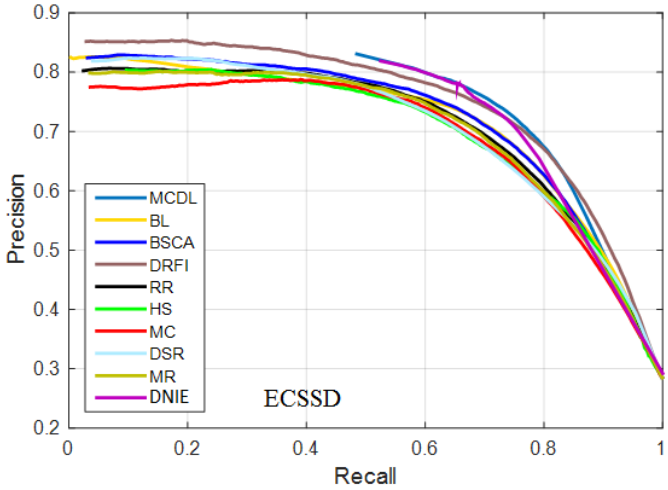


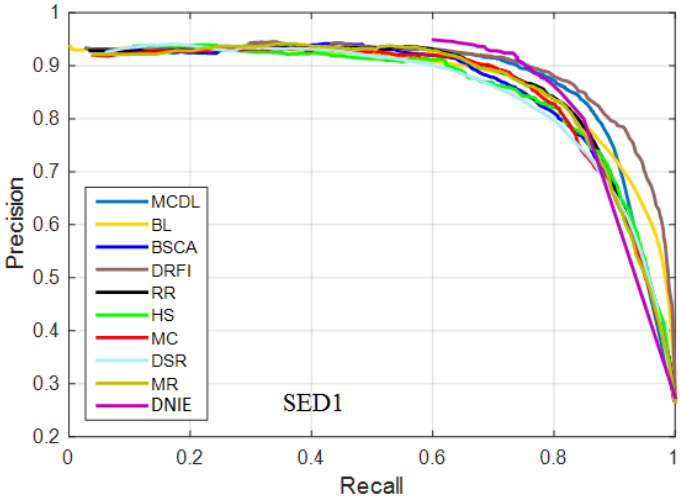
Figure 5.2 Saliency object detection results of different methods. From top to bottom: original image, our proposed DNIE, DNN based approach (MCDL [45]) and low-cues based approach (MR [14]).

We evaluate our proposed DNIE algorithm against nine state-of-the-art methods, including MCDL [45], DRFI [150], BL [15], MC [17], MR [14], RR [13], HS [22], BSCA [12] and DSR [151] on PASCAL-S [23], ECSSD [22], SED1 [139], SED2 [139] and DUT-OMRON [14] datasets respectively. These benchmark datasets have been introduced in the chapter 3.3.

The PR-curves of our DNIE method and state-of-the-art methods are drawn in Figure 5.3. According to PR-curves, DNIE can achieve the high performance on PASCAL-S, ECSSD and SED1 datasets, which is competitive to the top-1 method among the chosen comparisons. On SED2 and DUT-OMRON datasets, DNIE also favorably ranks top-3 among state-of-the-art counterparts.


Precision
Recall
MCDL
BL
BSCA
DRFI
RR
HS
MC
DSR
MR
DNIE
PASCAL-S


Precision
Recall
MCDL
BL
BSCA
DRFI
RR
HS
MC
DSR
MR
DNIE
ECSSD


Precision
Recall
MCDL
BL
BSCA
DRFI
RR
HS
MC
DSR
MR
DNIE
SED1

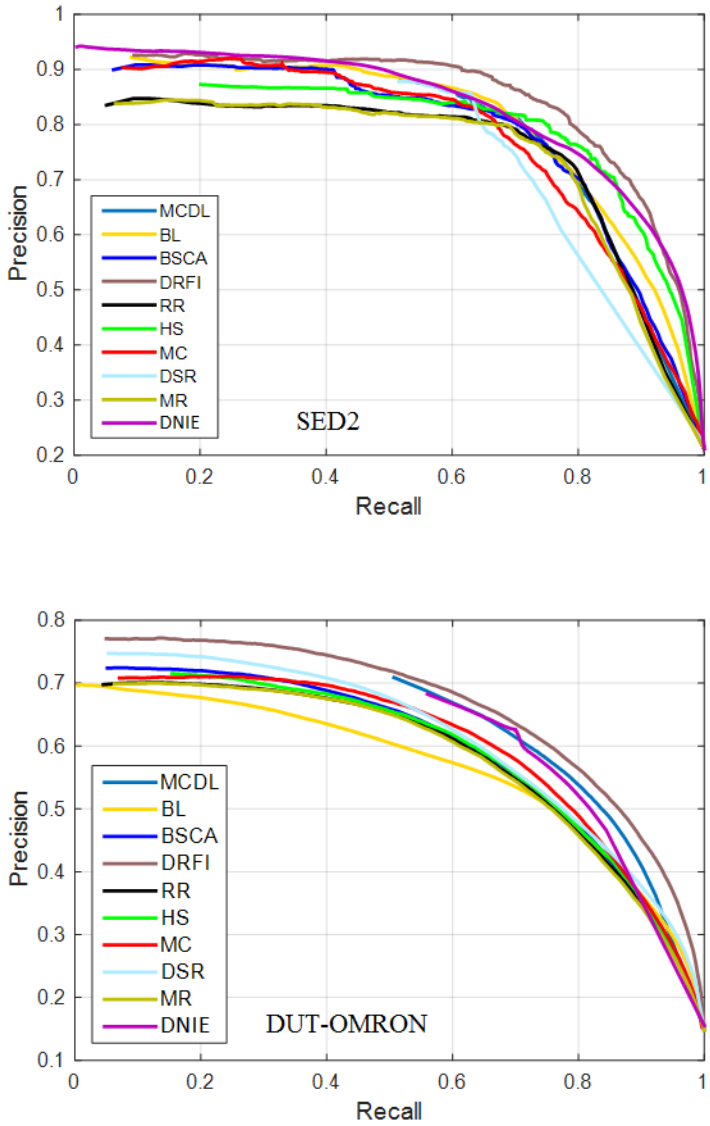


Figure 5.3 PR curve of benchmarking methods on five datasets.

Table 5.1 summaries F-measure of our DNIE method and state-of-the-art counterparts. F-measure can evaluate the comprehensive performance of a certain algorithm. The proposed DNIE beats the other methods on PASCAL-S, ECSSD and SED1 datasets, and achieves the second best results on DUT-OMRON dataset, which demonstrates that DNIE is robust across the different datasets.

Table 5.1 F-measure of benchmarking methods on five datasets. The best and second best results are shown in red and blue.

	PASCAL-S	ECSSD	SED1	SED2	DUT-OMRON
DNIE	0.7112	0.7479	0.8712	0.7904	0.7847
MCDL	0.6998	0.7469	0.8581	0.7847	0.6509
BL	0.6228	0.7161	0.8404	0.7934	0.5798
BSCA	0.6694	0.7180	0.8319	0.7797	0.6171
DRFI	0.6938	0.7358	0.8638	0.8226	0.6640
RR	0.6388	0.7097	0.8429	0.7692	0.6127
HS	0.6451	0.6975	0.8246	0.7815	0.6161
MC	0.6675	0.7028	0.8442	0.7755	0.6273
DSR	0.6506	0.6986	0.8186	0.7868	0.6269
MR	0.6188	0.7076	0.8410	0.7705	0.6108

Table 5.2 summaries MAE of our DNIE method and state-of-the-art counterparts. DNIE ranks top-2 on the five benchmarking datasets.

Table 5.2 MAE of benchmarking methods on five datasets. The best and second best results are shown in red and blue.

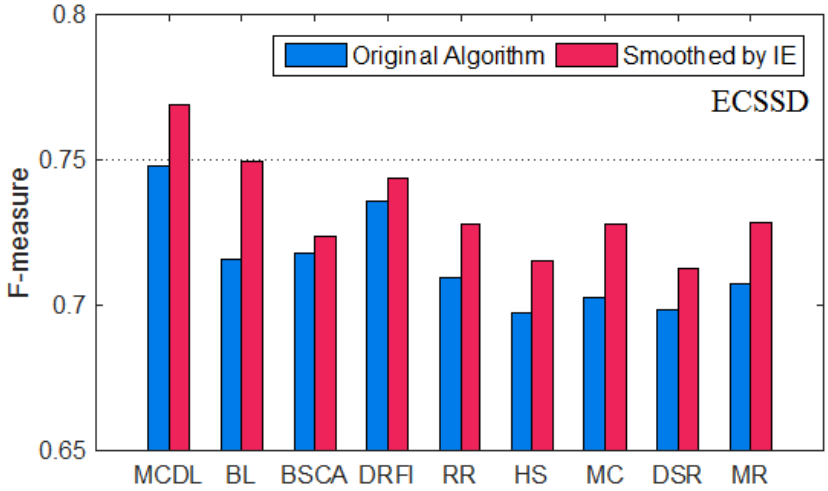
	PASCAL-S	ECSSD	SED1	SED2	DUT-OMRON
DNIE	0.1580	0.1790	0.0869	0.1217	0.1082
MCDL	0.1597	0.1752	0.0875	0.1074	0.1183
BL	0.2493	0.2620	0.1900	0.1403	0.2401
BSCA	0.2238	0.2235	0.1548	0.1583	0.1908
DRFI	0.2098	0.2256	0.1485	0.1403	0.1496
RR	0.2316	0.2235	0.1409	0.1614	0.1845
HS	0.2637	0.2686	0.1632	0.1951	0.2274
MC	0.2317	0.2513	0.1645	0.1804	0.1863
DSR	0.2079	0.2263	0.1599	0.1894	0.1388
MR	0.2588	0.2358	0.1431	0.1639	0.1868

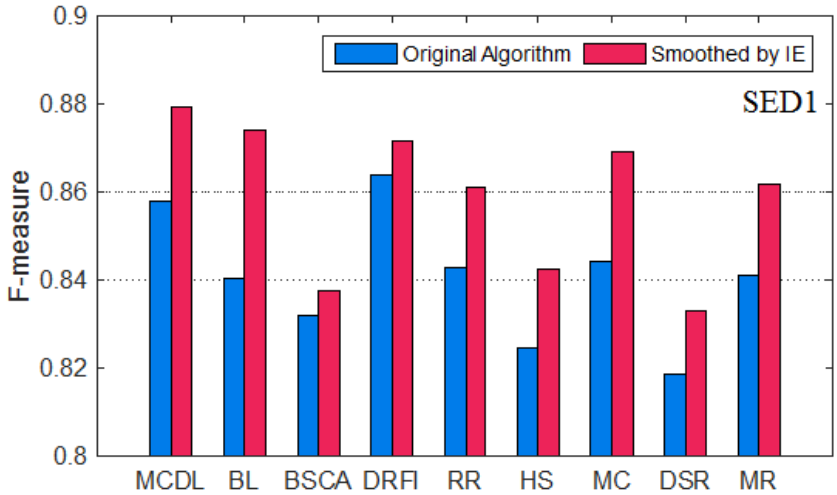
5.2.2. Evaluation on smooth penalty

As any other types of data penalty can be adopted to formulate our proposed image energy (IE), to evaluate our proposed saliency smooth penalty and its robustness, we further use the saliency maps generated by comparison methods to estimate the data penalty of IE. We then minimize the formulated IE and form the corresponding saliency maps as described in Chapter 5.1.4.

The results of F-measure in Figure 5.4 prove favorable improvements for all the comparison methods by 0.56% to 3.35% on SED1, 0.42% to 2.31% on SED2 and

0.09% to 3.50% on ECSSD. The detailed improvements for each comparison methods are shown in Figure 5.4.





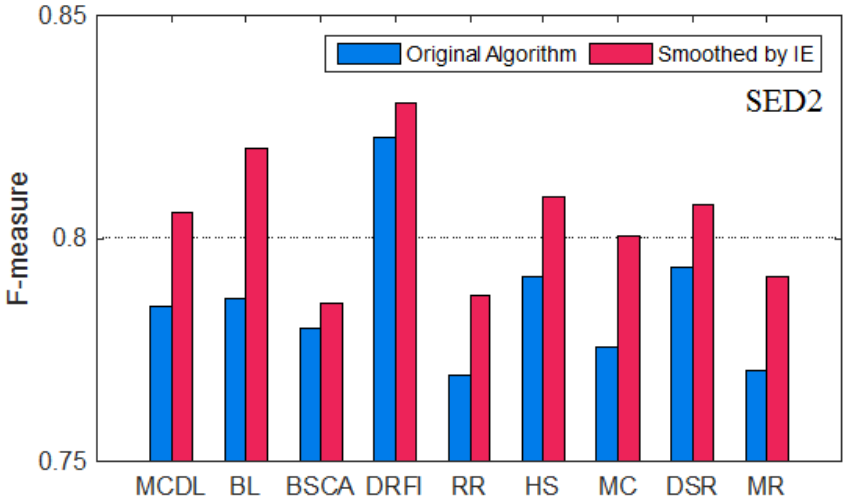


Figure 5.4 Quantitative improvements of state-of-the-art methods by our proposed image energy (IE).

As our proposed saliency image energy can boost the performance of existing methods, it can be utilized as the post-processing for other saliency detection methods which leads to smoother inner regions and more distinct region boundaries. For example, as shown in the starfish image of Figure 5.5, while some regions are wrongly labelled as foreground by multi-contexts combined DNN, the proposed image energy can well-suppress them. Since some parts of starfish are not captured by MCDL method, the results are improved by the image energy. Although MR method satisfactorily recognizes the starfish, the image energy enhances the saliency region so that decreases the mean absolute error.

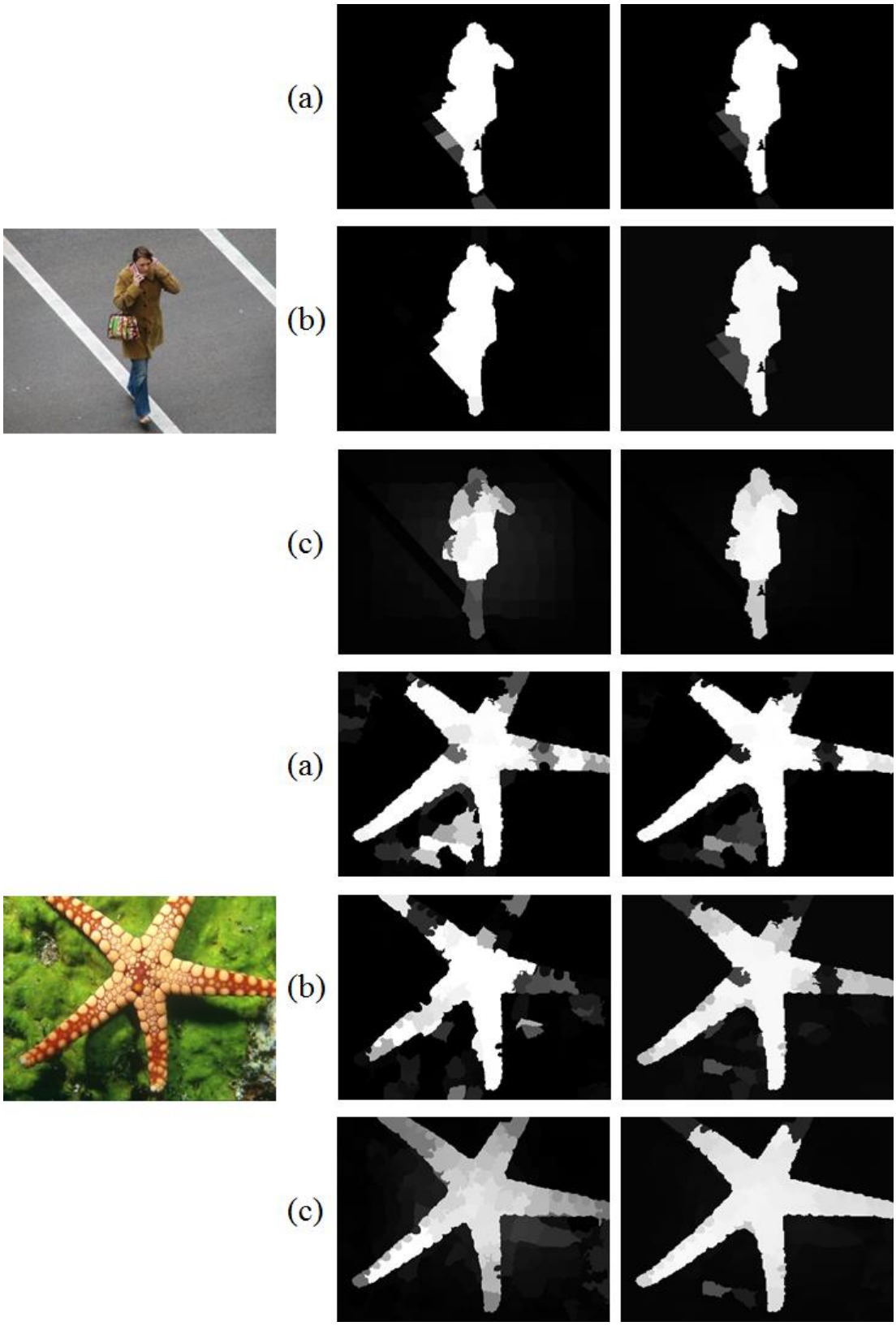


Figure 5.5 Examples of the improvements by proposed saliency image energy. From left to right: original images; original saliency maps by (a) multi-contexts combined DNN, (b) MCDL [45], (c) MR [14]; smoothed saliency maps after image energy minimization.

5.2.3. Comparison with other image energy

Carreira and Sminchisescu [158] proposed a novel image energy method (CMCP) for object segmentation. In 2014, Li et al. [23] applied CMCP algorithm to segment objects from a given eye fixation map produced by GBVS [36]. We performed such CMCP+GBVS on PASCAL-S dataset and measure the corresponding F-measure and MAE. The comparisons of CMCP+GBVS and our proposed DNIE method are listed in Table 5.3.

Table 5.3 F-measure and MAE of CMCP+GBVS and DNIE on PASCAL-S dataset.

	CMCP+GBVS	DNIE
F-measure	0.7111	0.7112
MAE	0.2130	0.1580

Although our method and GBVS+CMCP almost achieve the same F-measure (0.7112 vs 0.7111), our method strongly beats GBVS+CMCP in terms of MAE (0.1580 vs 0.2130). This is attributed to the following two reasons:

(a) CPMC measures the smooth penalty merely among adjacent pixels, while our work treats the image as a complete graph in superpixel scale, enabling smooth penalty to be measured in a holistic way;

(b) An inherent limitation of complete graph may lead to trivial errors. We therefore use region priors to guide the construction of the smooth penalty.

5.3. Summary

In this chapter, we have proposed a novel image energy with deep neural network to recognize the saliency object. An image segmentation approach is adopted to generate region-prior for image energy formulation. The saliency map can be eventually calculated by image energy minimization. In the experiments, we have evaluated our approach in comparison with nine state-of-the-art methods on five benchmark datasets. The experimental results show that our proposed approach favorably outperform the comparison methods. Furthermore, we have constructed the region-prior-based image energy with the data penalty measured by the results of comparison methods to evaluate the smooth penalty. The significant improvement of comparison methods prove an effective post-process served by our proposed image energy.

6. Discussion, conclusion and future work

6.1. Discussion

We discuss the overall performance of the three proposed object recognition models by directly comparing them over the same datasets. According to the experimental results in chapter 3, chapter 4 and chapter 5, the quantitative comparisons of the proposed BSFE, MCDN and DNIE methods can be summarized in Table 6.1 and Table 6.2.

Table 6.1 F-measure of BSFE, MCDN and DNIE on saliency detection datasets. The best and second best results are shown in red and blue.

	PASCAL-S	ECSSD	SED1	SED2
BSFE	0.6699	0.7080	0.8137	0.7815
MDCN	0.7041	0.7430	0.8619	0.7575
DNIE	0.7112	0.7479	0.8712	0.7904

Table 6.2 MAE of BSFE, MCDN and DNIE on saliency detection datasets. The best and second best results are shown in red and blue.

	PASCAL-S	ECSSD	SED1	SED2
BSFE	0.1926	0.2046	0.1132	0.1374
MDCN	0.1625	0.1813	0.0911	0.1279
DNIE	0.1580	0.1790	0.0869	0.1217

MCDN significantly outperforms BSFE in terms of F-measure and MAE. The superiority of MCDN can be attributed to the following two factors. Firstly, MCDN conducts convolution processing over the hidden layers, thus captures more detailed structures of input data; secondly, while BSFE requires vectorization input to SAE which may loss spatial information, MCDN directly adopt original multi-scaled images to DNN so that the spatial information can be feed-forwarded across the deep neural networks. While MDCN estimates the recognition maps superpixel by superpixel, DNIE refines the recognitions in global views by proposed saliency image energy. As shown in Figure 6.1, as the improvement of MDCN, DNIE can boost the recognition results by smoothing the inner regions and suppressing the outer backgrounds.

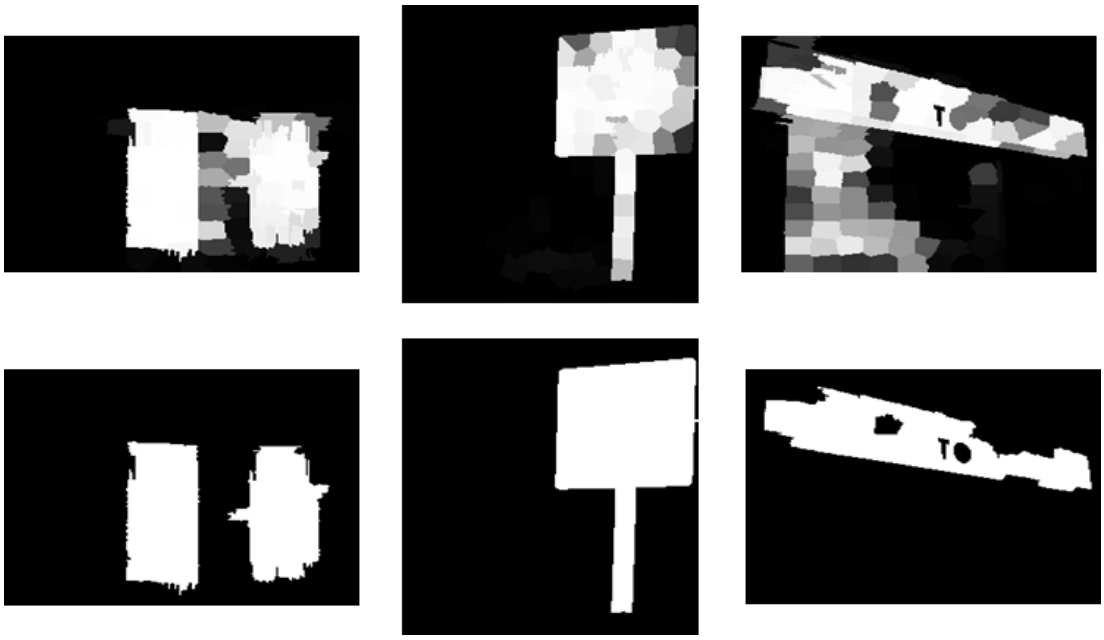


Figure 6.1 Example results of BSFE, MCDN and DNIE for saliency detection. From top to bottom: original images, saliency maps produced by BSFE, MCDN and DNIE.

Although our proposed MCDN and DNIE methods beat the state-of-the-art counterparts, they cannot generate more precise recognition maps in the case of complex foregrounds and multi-objects. For examples, as shown in the left column of Figure 5.2, our method fails to capture the windows of the bus as the main body and windows of the bus share much differential features. The semantic segmentation may address this challenge, in which the model is specially trained for the purpose of bus segmentation. As shown in the right column of Figure 4.3, some target objects are abandoned by our proposed algorithm when over two objects appear in the image. The main reason of this case is that the deep networks are trained by single-object image sets, thus it is not well adaptive to multi-objects images.

6.2. Conclusion

In this thesis, we proposed three object recognition models on multimodality images to solve the recognition challenges.

Firstly, we proposed a new comprehensive autoencoder for prostate recognition, followed by an image minimization scheme for refinement. Different from the most existing works with autoencoder, we let autoencoder itself serve as a classifier to focus on the prostate feature extraction, and the impacts by the irregular and complex background can be thus decreased. The comparative experiments with three classic classifiers and one atlas-based seeds-selection demonstrated the significant superiority of our proposed model for prostate recognition. We then applied the model on saliency object detection, and also achieved favorable performance on public datasets.

Secondly, in order to solve the challenges by complex imaging scenarios, we employed deep neural networks for feature extraction. As deep neural networks are invented on the basis of human brain machanism and contain more than tens of thousands parameters, the deep networks features can semantically and cognitively represent the intrinsic data structures of input data. Different from other multi-scaled deep networks methods, we proposed a uniform model to extract local and global features, thus do not require handcrafted combination of multi-scaled results. We validated this multi-contexts combined deep neural networks model for saliency object detection and prostate recognition. The favourable experiments results showed that our proposed object recognition model is effective and robust both on natural and biomedical images.

Thirdly, we designed a novel saliency image energy for the aim of more precise saliency object detection. To make the model more suitable for saliency detection, we imposed region priors on the image energy, on the basis of our three observations. Then we proposed a new saliency detection algorithm via integrating the saliency image energy and multi-contexts combined deep neural networks model.

The proposed new algorithm was compared with current state-of-the-art saliency detection methods on five well-recognized datasets. The experimental results showed that our algorithm can gain accurate and robust saliency recognitions. We further evaluated our proposed saliency energy model individually and demonstrated that it can be a post-process and refinement for most existing approaches.

6.3. Future work

In the future, we will investigate the algorithms of cancer detection and tumor staging with the recognized tissue and organ. Compared to tissue and organ recognition, the image noise and intensity inhomogeneity on biomedical images pose more significant challenges for cancer detection and tumor staging, as the speck may be labelled as false-positive cancer. Although deep neural networks may also be employed to address this issue, it still needs necessary improvements for accurate and robust results. Employing prior knowledge can be used to boost the performance of current deep neural networks, under the limitation of computation and memory capacities. For example, fully convolutional networks [44] can be applied on saliency object detection to generate dense pixel-wise recognition maps. However, the object boundaries cannot be precisely delineated as the detailed local contexts are gradually decayed during feedforward in deep neural networks. In this case, some segmentation priors (such as superpixel used in [43]) should be introduced to deep neural networks so that the results are encouraged to evolve into desirables. However, for cancer detection and tumor staging, the proper ways for such integration and the satisfactory balances between handcrafted priors and (un)supervised learning are still remained to be further studied.

7. References

[1] K. Yan, C. Li, X. Wang, A. Li, Y. Yuan, J. Kim, *et al.*, "Adaptive Background Search and Foreground Estimation for Saliency Detection via Comprehensive Autoencoder," presented at the 2016 The International Conference on Image Processing (ICIP), 2016.

[2] K. Yan, C. Li, X. Wang, Y. Yuan, A. Li, J. Kim, *et al.*, "Comprehensive autoencoder for prostate recognition on MR images," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1190-1194.

[3] J. Marin, D. Vázquez, A. Lopez, J. Amores, and B. Leibe, "Random forests of local experts for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2592-2599.

[4] X. Zhen, L. Shao, and X. Li, "Action recognition by spatio-temporal oriented energies," *Information Sciences*, vol. 281, pp. 295-309, 2014.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 1254-1259, 1998.

[6] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," in *ACM Transactions on graphics (TOG)*, 2007, p. 10.

[7] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: internet image montage," *ACM Transactions on Graphics (TOG)*, vol. 28, p. 124, 2009.

[8] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep Networks for Saliency Detection via Local Estimation and Global Search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183-3192.

[9] W. Qiu, J. Yuan, E. Ukwatta, Y. Sun, M. Rajchl, and A. Fenster, "Prostate segmentation: An efficient convex optimization approach with axial symmetry using 3-D TRUS and MR images," *Medical Imaging, IEEE Transactions on*, vol. 33, pp. 947-960, 2014.

[10] M.-P. Jolly, "Automatic segmentation of the left ventricle in cardiac MR and CT images," *International Journal of Computer Vision*, vol. 70, pp. 151-163, 2006.

[11] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of Alzheimer's disease with deep learning," in *Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on*, 2014, pp. 1015-1018.

[12] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency Detection via Cellular Automata," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 110-119.

[13] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. D. Feng, "Robust Saliency Detection via Regularized Random Walks Ranking," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 34, p. 2274, 2015.

[14] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 3166-3173.

[15] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient Object Detection via Bootstrap Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1884-1892.

[16] Y. Liu, Q. Cai, X. Zhu, J. Cao, and H. Li, "Saliency detection using two-stage scoring," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 4062-4066.

[17] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 1665-1672.

[18] K. Fu, C. Gong, I. Y. Gu, J. Yang, and P. Shi, "Salient object detection using normalized cut and geodesics," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 1100-1104.

[19] R. S. Srivatsa and R. V. Babu, "Salient object detection via objectness measure," in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 4481-4485.

[20] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior based salient object detection via deep reconstruction residual," 2014.

[21] M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, pp. 569-582, 2015.

[22] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 1155-1162.

[23] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 280-287.

[24] A. Li, "Medical image segmentation based on Dirichlet energies and priors," 2014.

[25] C. Y. Li, "Statistical Analysis Based Segmentation for Multimodality Images," University of Sydney, 2012.

[26] R. Grzeszick, L. Rothacker, and G. A. Fink, "Bag-of-features representations using spatial visual vocabularies for object classification," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, 2013, pp. 2867-2871.

[27] M. Alberti, J. Folkesson, and P. Jensfelt, "Relational approaches for joint object classification and scene similarity measurement in indoor environments," in *AAAI 2014 Spring Symposia: Qualitative Representations for Robots*, 2014.

[28] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 2008, pp. 1-4.

[29] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Computer Vision–ECCV 2010*, ed: Springer, 2010, pp. 366-379.

[30] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and whistles," in *Computer Vision–ECCV 2014*, ed: Springer, 2014, pp. 720-735.

[31] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, pp. 743-761, 2012.

[32] D. Kumar, A. Wong, and D. A. Clausi, "Lung Nodule Classification Using Deep Features in CT Images," in *Computer and Robot Vision (CRV), 2015 12th Conference on*, 2015, pp. 133-138.

[33] C. Lartizien, M. Rogez, E. Niaf, and F. Ricard, "Computer-aided staging of lymphoma patients with FDG PET/CT imaging based on textural information," *Biomedical and Health Informatics, IEEE Journal of*, vol. 18, pp. 946-955, 2014.

[34] A. Bianchi, J. V. Miller, E. T. Tan, and A. Montillo, "Brain tumor segmentation with symmetric texture and symmetric intensity-based decision forests," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, 2013, pp. 748-751.

[35] S. Ghose, A. Oliver, R. Martí, X. Lladó, J. C. Vilanova, J. Freixenet, *et al.*, "A survey of prostate segmentation methodologies in ultrasound, magnetic resonance and computed tomography images," *Computer methods and programs in biomedicine*, vol. 108, pp. 262-287, 2012.

[36] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545-552.

[37] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf, "Ranking on data manifolds," *Advances in neural information processing systems*, vol. 16, pp. 169-176, 2004.

[38] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, *et al.*, "Saliency Propagation From Simple to Difficult," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2531-2539.

[39] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, pp. 989-1005, 2009.

[40] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, *et al.*, "Learning to detect a salient object," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, pp. 353-367, 2011.

[41] J. Yang and M.-H. Yang, "Top-down visual saliency via joint crf and dictionary learning," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2296-2303.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[43] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep Image Saliency Computing via Progressive Representation Learning," 2015.

[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[45] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265-1274.

[46] W. Zou and N. Komodakis, "HARF: Hierarchy-Associated Rich Features for Salient Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 406-414.

[47] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 33, pp. 898-916, 2011.

[48] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 438-445.

[49] A. Li, C. Li, X. Wang, S. Eberl, D. D. Feng, and M. Fulham, "Automated Segmentation of Prostate MR Images Using Prior Knowledge Enhanced Random Walker," in *Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on*, 2013, pp. 1-7.

[50] S. Martin, V. Daanen, and J. Troccaz, "Atlas-based prostate segmentation using an hybrid registration," *International Journal of Computer Assisted Radiology and Surgery,* vol. 3, pp. 485-492, 2008.

[51] M. Samiee, G. Thomas, and R. Fazel-Rezai, "Semi-automatic prostate segmentation of MR images based on flow orientation," in *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, 2006, pp. 203-207.

[52] R. Zwiggelaar, Y. Zhu, and S. Williams, "Semi-automatic segmentation of the prostate," in *Pattern Recognition and Image Analysis*, ed: Springer, 2003, pp. 1108-1116.

[53] D. Flores-Tapia, G. Thomas, N. Venugopa, B. McCurdy, and S. Pistorius, "Semi automatic MRI prostate segmentation based on wavelet multiscale products," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, 2008, pp. 3020-3023.

[54] C. Li, X. Wang, Y. Xia, S. Eberl, Y. Yin, and D. D. Feng, "Automated PET-guided liver segmentation from low-contrast CT volumes using probabilistic atlas," *Computer methods and programs in biomedicine,* vol. 107, pp. 164-174, 2012.

[55] D. Terzopoulos, "On matching deformable models to images," in *Topical Meeting on Machine Vision Tech. Digest Series*, 1987, pp. 160-167.

[56] S. Martin, J. Troccaz, and V. Daanen, "Automated segmentation of the prostate in 3D MR images using a probabilistic atlas and a spatially constrained deformable model," *Medical physics,* vol. 37, pp. 1579-1590, 2010.

[57] M. Kirschner, F. Jung, and S. Wesarg, "Automatic prostate segmentation in MR images with a probabilistic active shape model," *MICCAI Grand Challenge: Prostate MR Image Segmentation,* vol. 2012, 2012.

[58] Y. Guo, Y. Gao, Y. Shao, T. Price, A. Oto, and D. Shen, "Deformable segmentation of 3D MR prostate images via distributed discriminative dictionary and ensemble learning," *Medical physics,* vol. 41, p. 072303, 2014.

[59] Y. Zhan and D. Shen, "Deformable segmentation of 3-D ultrasound prostate images using statistical texture matching method," *Medical Imaging, IEEE Transactions on,* vol. 25, pp. 256-272, 2006.

[60] R. Toth, P. Tiwari, M. Rosen, G. Reed, J. Kurhanewicz, A. Kalyanpur, *et al.*, "A magnetic resonance spectroscopy driven initialization scheme for active shape model based prostate segmentation," *Medical Image Analysis,* vol. 15, pp. 214-225, 2011.

[61] S. S. Chandra, J. A. Dowling, K.-K. Shen, P. Raniga, J. P. Pluim, P. B. Greer, *et al.*, "Patient specific prostate segmentation in 3-D magnetic resonance

images," *Medical Imaging, IEEE Transactions on,* vol. 31, pp. 1955-1964, 2012.

[62] B. Maan and F. van der Heijden, "Prostate MR image segmentation using 3D active appearance models," 2012.

[63] C. Xu, D. L. Pham, and J. L. Prince, "Image segmentation using deformable models," *Handbook of medical imaging,* vol. 2, pp. 129-174, 2000.

[64] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding,* vol. 61, pp. 38-59, 1995.

[65] R. Toth, B. N. Bloch, E. M. Genega, N. M. Rofsky, R. E. Lenkinski, M. A. Rosen, *et al.,* "Accurate prostate volume estimation using multifeature active shape models on T2-weighted MRI," *Academic radiology,* vol. 18, pp. 745-754, 2011.

[66] X. Tang, Y. Jeong, R. J. Radke, and G. T. Chen, "Geometric-model-based segmentation of the prostate and surrounding structures for image-guided radiotherapy," in *Electronic Imaging 2004,* 2004, pp. 168-176.

[67] Y. Zhu, S. Williams, and R. Zwiggelaar, "A hybrid ASM approach for sparse volumetric data segmentation," *Pattern Recognition and Image Analysis,* vol. 17, pp. 252-258, 2007.

[68] A. C. Hodge, A. Fenster, D. B. Downey, and H. M. Ladak, "Prostate boundary segmentation from ultrasound images using 2D active shape models: Optimisation and extension to 3D," *Computer methods and programs in biomedicine,* vol. 84, pp. 99-113, 2006.

[69] Q. Feng, M. Foskey, W. Chen, and D. Shen, "Segmenting CT prostate images using population and patient-specific statistics for radiotherapy," *Medical Physics,* vol. 37, pp. 4121-4132, 2010.

[70] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "The use of active shape models for locating structures in medical images," in *Information Processing in Medical Imaging,* 1993, pp. 33-47.

[71] D. Shen, Y. Zhan, and C. Davatzikos, "Segmentation of prostate boundaries from ultrasound images using statistical shape model," *Medical Imaging, IEEE Transactions on,* vol. 22, pp. 539-551, 2003.

[72] R. Toth and A. Madabhushi, "Multifeature landmark-free active appearance models: application to prostate MRI segmentation," *Medical Imaging, IEEE Transactions on,* vol. 31, pp. 1638-1650, 2012.

[73] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence,* pp. 681-685, 2001.

[74] S. Fan, L. K. Voon, and N. W. Sing, "3D prostate surface detection from ultrasound images based on level set method," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2002,* ed: Springer, 2002, pp. 389-396.

[75] K. Zhang, L. Zhang, H. Song, and W. Zhou, "Active contours with selective local or global segmentation: a new formulation and level set method," *Image and Vision computing,* vol. 28, pp. 668-676, 2010.

[76] W. Xiong, A. L. Li, S. H. Ong, and Y. Sun, "Automatic 3D prostate MR image segmentation using graph cuts and level sets with shape prior," in *Advances in Multimedia Information Processing–PCM 2013,* ed: Springer, 2013, pp. 211-220.

[77] N. N. Kachouie, P. Fieguth, and S. Rahnamayan, "An elliptical level set method for automatic TRUS prostate image segmentation," in *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, 2006, pp. 191-196.

[78] C. Li, R. Huang, Z. Ding, J. C. Gatenby, D. N. Metaxas, and J. C. Gore, "A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI," *Image Processing, IEEE Transactions on*, vol. 20, pp. 2007-2016, 2011.

[79] A. Tsai, A. Yezzi Jr, W. Wells, C. Tempany, D. Tucker, A. Fan, *et al.*, "A shape-based approach to the segmentation of medical imagery using level sets," *Medical Imaging, IEEE Transactions on*, vol. 22, pp. 137-154, 2003.

[80] A. Zaim and J. Jankun, "An energy-based segmentation of prostate from ultrasouind images using dot-pattern select cells," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, pp. I-297-I-300.

[81] C. Knoll, M. Alcañiz, V. Grau, C. Monserrat, and M. C. Juan, "Outlining of the prostate using snakes with shape restrictions based on the wavelet transform (Doctoral Thesis: Dissertation)," *Pattern Recognition*, vol. 32, pp. 1767-1781, 1999.

[82] H. M. Ladak, F. Mao, Y. Wang, D. B. D. B. Downey, D. Steinman, and A. Fenster, "Prostate segmentation from 2d ultrasound images," in *Engineering in Medicine and Biology Society, 2000. Proceedings of the 22nd Annual International Conference of the IEEE*, 2000, pp. 3188-3191.

[83] M. Ding, C. Chen, Y. Wang, I. Gyacskov, and A. Fenster, "Prostate segmentation in 3D US images using the cardinal-spline-based discrete dynamic contour," in *Medical Imaging 2003*, 2003, pp. 69-76.

[84] A. Jendoubi, J. Zeng, and M. F. Chouikha, "Segmentation of prostate ultrasound images using an improved snakes model," in *Signal Processing, 2004. Proceedings. ICSP'04. 2004 7th International Conference on*, 2004, pp. 2568-2571.

[85] P. A. Yushkevich, J. Piven, H. C. Hazlett, R. G. Smith, S. Ho, J. C. Gee, *et al.*, "User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability," *Neuroimage*, vol. 31, pp. 1116-1128, 2006.

[86] W. Du, S. Wang, A. Oto, and Y. Peng, "Graph-based prostate extraction in T2-weighted images for prostate cancer detection," in *Fuzzy Systems and Knowledge Discovery (FSKD), 2015 12th International Conference on*, 2015, pp. 1225-1229.

[87] S. Ghose, J. Mitra, A. Oliver, R. Marti, X. Llado, J. Freixenet, *et al.*, "Graph cut energy minimization in a probabilistic learning framework for 3D prostate segmentation in MRI," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2012, pp. 125-128.

[88] A. S. Korsager, V. Fortunati, F. van der Lijn, J. Carl, W. Niessen, L. R. Østergaard, *et al.*, "The use of atlas registration and graph cuts for prostate segmentation in magnetic resonance images," *Medical physics*, vol. 42, pp. 1614-1624, 2015.

[89] D. Mahapatra, "Graph cut based automatic prostate segmentation using learned semantic information," in *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, 2013, pp. 1316-1319.

[90] D. Mahapatra and J. M. Buhmann, "Visual saliency based active learning for prostate mri segmentation," in *Machine Learning in Medical Imaging*, ed: Springer, 2015, pp. 9-16.

[91] Q. Song, X. Wu, Y. Liu, M. Smith, J. Buatti, and M. Sonka, "Optimal graph search segmentation using arc-weighted graph for simultaneous surface detection of bladder and prostate," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, ed: Springer, 2009, pp. 827-835.

[92] Z. Tian, L. Liu, and B. Fei, "A supervoxel-based segmentation for prostate MR images," in *SPIE Medical Imaging*, 2015, pp. 941318-941318-7.

[93] K. Wu, C. Garnier, H. Shu, and J.-L. Dillenseger, "Prostate segmentation on T2 MRI using Optimal Surface Detection," *IRBM*, vol. 34, pp. 287-290, 2013.

[94] M. Zouqi and J. Samarabandu, "Prostate segmentation from 2-D ultrasound images using graph cuts and domain knowledge," in *Computer and Robot Vision, 2008. CRV'08. Canadian Conference on*, 2008, pp. 359-362.

[95] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 1124-1137, 2004.

[96] Y. Artan, M. A. Haider, and I. S. Yetik, "Prostate cancer segmentation using multispectral random walks," in *Prostate Cancer Imaging. Computer-Aided Diagnosis, Prognosis, and Intervention*, ed: Springer, 2010, pp. 15-24.

[97] P. Khurd, L. Grady, K. Gajera, M. Diallo, P. Gall, M. Requardt, *et al.*, "Facilitating 3d spectroscopic imaging through automatic prostate localization in mr images using random walker segmentation initialized via boosted classifiers," in *Prostate Cancer Imaging. Image Analysis and Image-Guided Interventions*, ed: Springer, 2011, pp. 47-56.

[98] L. Grady, "Random walks for image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, pp. 1768-1783, 2006.

[99] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144-152.

[100] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, pp. 278-282.

[101] S. Ghose, A. Oliver, J. Mitra, R. Martí, X. Lladó, J. Freixenet, *et al.*, "A supervised learning framework of statistical shape and probability priors for automatic prostate segmentation in ultrasound images," *Medical Image Analysis*, vol. 17, pp. 587-600, 2013.

[102] S. Ghose, J. Mitra, A. Oliver, R. Martí, X. Lladó, J. Freixenet, *et al.*, "A random forest based classification approach to prostate segmentation in MRI," *MICCAI Grand Challenge: Prostate MR Image Segmentation*, vol. 2012, 2012.

[103] D. Mahapatra and J. M. Buhmann, "Prostate mri segmentation using learned semantic knowledge and graph cuts," *Biomedical Engineering, IEEE Transactions on*, vol. 61, pp. 756-764, 2014.

[104] E. Moschidis and J. Graham, "Automatic differential segmentation of the prostate in 3-D MRI using Random Forest classification and graph-cuts optimization," in *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, 2012, pp. 1727-1730.

[105] Y. Gao, S. Liao, and D. Shen, "Prostate segmentation by sparse representation based classification," *Medical physics,* vol. 39, pp. 6372-6387, 2012.
[106] W. Li, S. Liao, Q. Feng, W. Chen, and D. Shen, "Learning image context for segmentation of the prostate in CT-guided radiotherapy," *Physics in medicine and biology,* vol. 57, p. 1283, 2012.
[107] P. Ghosh and M. Mitchell, "Segmentation of medical images using a genetic algorithm," in *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, 2006, pp. 1171-1178.
[108] Y. Guo, Y. Gao, and D. Shen, "Deformable MR Prostate Segmentation via Deep Feature Learning and Sparse Patch Matching," 2015.
[109] Z. Tian, L. Liu, Z. Zhang, and B. Fei, "Superpixel-based Segmentation for 3D Prostate MR Images," 2015.
[110] N. Makni, P. Puech, R. Lopes, A.-S. Dewalle, O. Colot, and N. Betrouni, "Combining a deformable model and a probabilistic framework for an automatic 3D segmentation of prostate on MRI," *International journal of computer assisted radiology and surgery,* vol. 4, pp. 181-188, 2009.
[111] L. Gong, S. D. Pathak, D. R. Haynor, P. S. Cho, and Y. Kim, "Parametric shape modeling using deformable superellipses for prostate segmentation," *Medical Imaging, IEEE Transactions on,* vol. 23, pp. 340-349, 2004.
[112] Q. Gao, A. Asthana, T. Tong, Y. Hu, D. Rueckert, and P. Edwards, "Hybrid Decision Forests for Prostate Segmentation in Multi-channel MR Images," in *2014 22nd International Conference on Pattern Recognition (ICPR)*, 2014, pp. 3298-3303.
[113] P. Wu, Y. Liu, Y. Li, and B. Liu, "Robust Prostate Segmentation Using Intrinsic Properties of TRUS Images," *Medical Imaging, IEEE Transactions on,* vol. 34, pp. 1321-1335, 2015.
[114] C. Lu, S. Chelikani, X. Papademetris, J. P. Knisely, M. F. Milosevic, Z. Chen, *et al.*, "An integrated approach to segmentation and nonrigid registration for application in image-guided pelvic radiotherapy," *Medical Image Analysis,* vol. 15, pp. 772-785, 2011.
[115] S. Liao, Y. Gao, A. Oto, and D. Shen, "Representation learning: A unified deep learning framework for automatic prostate MR segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*, ed: Springer, 2013, pp. 254-261.
[116] L. Lu, A. Barbu, M. Wolf, J. Liang, L. Bogoni, M. Salganicoff, *et al.*, "Simultaneous detection and registration for ileo-cecal valve detection in 3D CT colonography," in *Computer Vision–ECCV 2008*, ed: Springer, 2008, pp. 465-478.
[117] L. Lu, A. Barbu, M. Wolf, J. Liang, M. Salganicoff, and D. Comaniciu, "Accurate polyp segmentation for 3D CT colongraphy using multi-staged probabilistic binary learning and compositional model," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.
[118] H. Ling, S. K. Zhou, Y. Zheng, B. Georgescu, M. Suehling, and D. Comaniciu, "Hierarchical, learning-based automatic liver segmentation," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.
[119] L. Yang, B. Georgescu, Y. Zheng, P. Meer, and D. Comaniciu, "3D ultrasound tracking of the left ventricle using one-step forward prediction and

data fusion of collaborative trackers," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1-8.

[120] M. M. Center, A. Jemal, J. Lortet-Tieulent, E. Ward, J. Ferlay, O. Brawley, *et al.*, "International variation in prostate cancer incidence and mortality rates," *European urology*, vol. 61, pp. 1079-1092, 2012.

[121] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length, and Helmholtz free energy," *Advances in neural information processing systems*, pp. 3-3, 1994.

[122] Q. Li, W. Cai, and D. D. Feng, "Lung image patch classification with automatic feature learning," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 6079-6082.

[123] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, pp. 2274-2282, 2012.

[124] N. Wang, J. Melchior, and L. Wiskott, "An analysis of Gaussian-binary restricted Boltzmann machines for natural images," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2012, pp. 287-292.

[125] Y. Ouyang, W. Liu, W. Rong, and Z. Xiong, "Autoencoder-Based Collaborative Filtering," in *Neural Information Processing*, 2014, pp. 284-291.

[126] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-Stage Learning to Predict Human Eye Fixations via SDAEs," 2015.

[127] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," DTIC Document1985.

[128] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, pp. 1798-1828, 2013.

[129] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, *et al.*, "Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge," *Medical image analysis*, vol. 18, pp. 359-373, 2014.

[130] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *Image Processing, IEEE Transactions on*, vol. 24, pp. 5706-5722, 2015.

[131] S. Qureshi, *Embedded image processing on the TMS320C6000TM DSP: examples in code composer studioTM and MATLAB*: Springer Science & Business Media, 2005.

[132] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, *et al.*, "Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology images," 2015.

[133] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2368-2375.

[134] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs for salient object detection in images," *Image Processing, IEEE Transactions on*, vol. 19, pp. 3232-3242, 2010.

[135] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 22, pp. 888-905, 2000.
[136] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, *et al.*, "Theano: A CPU and GPU math compiler in Python," in *Proc. 9th Python in Science Conf*, 2010, pp. 1-7.
[137] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, *et al.*, "Theano: new features and speed improvements," *arXiv preprint arXiv:1211.5590,* 2012.
[138] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 853-860.
[139] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 34, pp. 315-327, 2012.
[140] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on*, 2009, pp. 1597-1604.
[141] S. He and R. W. Lau, "Saliency detection with flash and no-flash image pairs," in *Computer Vision–ECCV 2014*, ed: Springer, 2014, pp. 110-124.
[142] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Computer Vision–ECCV 2012*, ed: Springer, 2012, pp. 29-42.
[143] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 10-17.
[144] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision,* vol. 59, pp. 167-181, 2004.
[145] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," in *Computer vision–ECCV 2008*, ed: Springer, 2008, pp. 705-718.
[146] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807-814.
[147] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167,* 2015.
[148] J. Moody, S. Hanson, A. Krogh, and J. A. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems,* vol. 4, pp. 950-957, 1995.
[149] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for matlab," in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, 2015, pp. 689-692.
[150] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 2083-2090.
[151] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 2976-2983.

[152] O. Veksler, "Image segmentation by nested cuts," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 2000, pp. 339-344.

[153] H. Ishikawa and D. Geiger, "Segmentation by grouping junctions," in *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, 1998, pp. 125-131.

[154] D. Greig, B. Porteous, and A. H. Seheult, "Exact maximum a posteriori estimation for binary images," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 271-279, 1989.

[155] Y. Boykov and D. P. Huttenlocher, "A new bayesian framework for object recognition," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, 1999.

[156] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," in *ACM Transactions on Graphics (ToG)*, 2003, pp. 277-286.

[157] J. Yan, Y. Yu, X. Zhu, Z. Lei, and S. Z. Li, "Object Detection by Labeling Superpixels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5107-5116.

[158] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 3241-3248.