Research and Practice on Visual Attention Module Design for Saliency Detection

Wang Lu

Abstract 1
1 Introduction 4
1.1 Research background and motivation 4
1.2 Introduction to visual attention based saliency detection 5
1.3 Existing challenges 5
1.4 Thesis outline 6
2 Attention module design for saliency detection 7
2.1 Development of attention for saliency detection 7
2.1.1 Conventional semantic attention 8
2.1.2 Deep CNN based method 11
2.1.3 The development and requirement in industry 14
2.2 The proposed semantic attention for saliency detection 15
2.2.1 Research motivation 15
2.2.2 Semantic attention mechanism 17
2.2.2.1 Proposal selection 19
2.2.2.2 Proposal reweighting 23
2.2.3 Simulation 25
2.2.4 Conclusion 31
2.3 The proposed visual attention in Deep Learning 32
2.3.1 Research motivation 32
2.3.2 Overview of attention mechanism in Deep Learning 32
2.3.3 The proposed pixel level attention for segmentation 39
2.3.3.1 Pyramid Attention Atrous Spatial Pyramid Pooling (PA-ASPP) 40
2.3.3.2 Scale-aware selection (SAS) 42
2.3.4 The Proposed Deep Segment Network (DSNet) 43
2.3.5 Conclusion 44
3 Visual attention in medical imaging applications 45
3.1 AI-aided diagnosis for medical image analytics 45
3.1.1 Status of AI-aided diagnosis for medical image analysis 45
3.1.2 Major challenges 46
3.1.3 Previous works 48
3.2 Visual attention system for breast cancer detection 48
3.2.1 Background introduction of breast cancer detection 48
3.2.2 Deep Regional Metastases Segmentation (DRMS) framework 51
3.2.3 Patch-level metastasis localization 54

3.2.4 Slide-Level Metastases Detection 55
3.2.5 Patient-level Classification 56
3.2.6 Performance analysis 59
3.2.7 Conclusion 66
4 Overall conclusion and future work 68
4.1 Conclusion 68
4.2 Future work 69
Acknowledgement 70
Reference 71

Abstract

Visual attention is an important mechanism in the human visual system. When human observe images and videos, they usually do not describe all the contents in them. Instead, they tend to talk about the semantically important regions and objects in the images. The human eye is usually attracted by some regions of interest rather than the entire scene. These regions of interest that present the mainly meaningful or semantic content are called saliency region.

Visual saliency detection refers to the use of intelligent algorithms to simulate human visual attention mechanism, extract both the low-level features and high-level semantic information and localize the salient object regions in images and videos. The generated saliency map indicates the regions that are likely to attract human attention.

As a fundamental problem of image processing and computer vision, visual saliency detection algorithms have been extensively studied by researchers to solve practical tasks, such as image and video compression, image retargeting, object detection, etc. The visual attention mechanism adopted by saliency detection in general are divided into two categories, namely the bottom-up model and top-down model. The bottom-up attention algorithm focuses on utilizing the low-level visual features such as colour and edges to locate the salient objects. While the top-down attention utilizes the supervised learning to detect saliency.

In recent years, more and more research tend to design deep neural networks with attention mechanisms to improve the accuracy of saliency detection. The design of deep attention neural network is inspired by human visual attention. The main goal is to enable the network to automatically capture the information that is critical to the target tasks and suppress irrelevant information, shift the attention from focusing on all to local. Currently various domain's attention has been developed for saliency detection and semantic segmentation, such as the spatial attention module in convolution network, it generates a spatial attention map by utilizing the inter-spatial relationship of features; the channel attention module produces a attention by

exploring the inter-channel relationship of features. All these well-designed attentions have been proven to be effective in improving the accuracy of saliency detection.

This paper investigates the visual attention mechanism of salient object detection and applies it to digital histopathology image analysis for the detection and classification of breast cancer metastases. As shown in following contents, the main research contents include three parts:

First, we studied the semantic attention mechanism and proposed a semantic attention approach to accurately localize the salient objects in complex scenarios. The proposed semantic attention uses Faster-RCNN to capture high-level deep features and replaces the last layer of Faster-RCNN by a FC layer and sigmoid function for visual saliency detection; it calculates proposals' attention probabilities by comparing their feature distances with the possible salient object. The proposed method introduces a re-weighting mechanism to reduce the influence of the complexity background, and a proposal selection mechanism to remove the background noise to obtain objects with accurate shape and contour. The simulation result shows that the semantic attention mechanism is robust to images with complex background due to the consideration of high-level object concept, the algorithm achieved outstanding performance among the salient object detection algorithms in the same period.

Second, we designed a deep segmentation network (DSNet) for saliency object prediction. We explored a Pyramidal Attentional ASPP (PA-ASPP) module which can provide pixel level attention. DSNet extracts multi-level features with dilated ResNet-101 and the multiscale contextual information was locally weighted with the proposed PA-ASPP. The pyramid feature aggregation encodes the multi-level features from three different scales. This feature fusion incorporates neighboring scales of context features more precisely to produce better pixel-level attention. Finally, we use a scale-aware selection (SAS) module to locally weight multi-scale contextual features, capture important contexts of ASPP for the accurate and consistent dense prediction. The simulation results demonstrated that the proposed PA-ASPP is effective and can generate more coherent results. Besides, with the SAS, the model can adaptively capture the regions with different scales effectively.

Finally, based on previous research on attentional mechanisms, we proposed a novel Deep Regional Metastases Segmentation (DRMS) framework for the detection and classification of breast cancer metastases. As we know, the digitalized whole slide image has high-resolution, usually has gigapixel, however the size of abnormal region is often relatively small, and most of the slide region are normal. The highly trained pathologists usually localize the regions of interest first in the whole slide, then perform precise examination in the selected regions. Even though the process is time-consuming and prone to miss diagnosis. Through observation and analysis, we believe that visual attention should be perfectly suited for the application of digital pathology image analysis. The integrated framework for WSI analysis can capture the granularity and variability of WSI, rich information from multi-grained pathological image. We first utilize the proposed attention mechanism based DSNet to detect the regional metastases in patch-level. Then, adopt the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to predict the whole metastases from individual slides. Finally, determine patient-level pN-stages by aggregating each individual slide-level prediction. In combination with the above techniques, the framework can make better use of the multi-grained information in histological lymph node section of whole-slice images. Experiments on large-scale clinical datasets (e.g., CAMELYON17) demonstrate that our method delivers advanced performance and provides consistent and accurate metastasis detection.

Key words: saliency detection, semantic attention, self-attention, deep learning, feature extraction

1 Introduction

1.1 Research background and motivation

In recent years, with the development of Deep Learning (DL) technology, the digitization of information and the improvement of computing capabilities, Computer Vision (CV) technology has developed rapidly. As one of the fundamental and key technologies of CV, saliency detection has been widely used in various downstream research work. The visual attention mechanism studied in this paper is an indispensable technology for saliency detection and plays an important role in achieving the high accuracy and high efficiency. Below are some typical studies related to the saliency detection technology.

a) Various image and video coding technologies and standards have been developed to maintain high image quality at high compression rates. The basic idea is that the codec does not need to compress the entire frame with uniform quality. It can utilize the characteristics of the human visual system, extract features of human interest, such as texture, motion, shape, etc., to divide each frame of the video sequence into image regions of arbitrary shapes. Through saliency detection, use more bits to compress the saliency regions to ensure high quality, while fewer bits in other areas to reduce the overall bitrate. MPEG-4 is the first codec standard to adopt object-based coding.
b) Image retargeting aims at displaying images without distortion on media of various sizes (cell phones, projection screens). With saliency detection, the image can be resized by expanding and shrinking the noninformation regions, and maximum guaranteed the display of saliency area without distortion.
c) Object detection as one of the basic CV tasks has been widely used in different application fields, such as parts sorting in industrial scene, the image-based lesion detection in medical scenes and the scene analysis in security field. Fast and accurate detection has always been the goal pursued by researchers. Through the saliency detection, the object detection can only occur in the salient area most likely to contain the object, rather than the entire image, ensuring accuracy while reducing the amount of calculation.

1.2 Introduction to visual attention based saliency detection

When observing images and videos, the human eyes are usually attracted by some small regions of interest rather than the entire scene. These regions of interest present the main meaningful semantic content of the scene, usually the human brain will automatically process these regions, while selectively ignoring regions that are not of interest. This is the so-called human visual attention mechanism, in which the area of interest is called the salient region.

Visual saliency detection refers to the use of intelligent algorithms to simulate human visual attention mechanism, to localize and extract the most conspicuous and eye-attracting object regions in images and videos. The generated saliency map indicates regions that are likely to attract human attention.

Existing saliency detection algorithms in general are categorized into bottom-up models and top-down models. The bottom-up models are unsupervised, major focus on utilizing the low-level visual features such as colour and edges to locate the salient objects. While the top-down models utilize the supervised learning to detect saliency. In recent years, more and more saliency detection research has tilted toward deep learning based solution.

1.3 Existing challenges

Although many new algorithms have been proposed and the saliency detection accuracy has been greatly improved, there are still some challenges that need to be overcome. The first challenge is how to locate the salient object accurate, especially in complex background environment. The second challenge is that the algorithm ability to understand and utilize contextual information of the image is critical to achieving accurate saliency object segmentation. Increasing the network receptive field by consecutive down sampling is a straightforward way to capture contextual clues. However, the number of down-sampling operations performed is restricted, as performing down-sampling can cause spatial information loss. Third, the attention mechanism algorithm needs to be combined with other algorithms to form a systematic

solution before it can be used for practical applications. It is also a challenge for researchers to make the results of attention mechanism effective in practical applications. In this study, we try to address these challenges.

1.4 Thesis outline

This paper is organized as follows. The current chapter gives a brief introduction on saliency detection technology, application, and existing challenges. Chapter 2 first gives a detailed overview of the attention module design for saliency detection, then introduces the proposed deep feature based semantic attention design and the visual attention module of the deep learning model. Chapter 3, the AI-assisted pathological image analysis framework for breast cancer screening is introduced in detail, and the effect of attention mechanism on system performance is analysed. In chapter 4, conclusions and the future work are introduced.

2 Attention module design for saliency detection

Saliency detection in computer vision aims to find salient objects in images and it is usually called salient object detection (SOD). The goal of SOD is to identify the most visually distinctive objects or regions in an image and then segment them out from the background. Different from other segmentation-like tasks, such as semantic segmentation which clusters all the objects inside the image, salient object detection pays more attention to very few objects that are interesting, attractive, and important.

Since L. Itti et al. published [49] in 1998, the research on SOD has received widespread attention. Since then, taking 2014 as the boundary, SOD can be roughly divided into two eras: conventional methods and deep learning methods. The approaches adopted by conventional SOD usually assign high probability values to salient objects in a scene to produce the saliency map. Regional contrast has been frequently used in conventional SOD; various complementary heuristic saliency priors have also been deployed to capture the conspicuous object regions in images. These conventional methods have been proven to be efficient and effective in relatively simple scenes with a single object and/or clean background. However, in practice, objects of different shapes, sizes, and positions, low contrast between objects and backgrounds, and complex and variable backgrounds all pose challenges for SOD prediction. The rapidly evolving deep learning-based techniques in computer vision have shown powerful capabilities in solving these problems.

SOD has shown potential in a various application field, such as the real-time security surveillance, the autonomous vehicles, and medical image processing. Due to the different visual appearances and semantic components, the attention mechanisms in these applications are quite different. In our study, we first focus the attention module design on natural image processing, and then delves further into the field of medical image processing.

2.1 Development of attention for saliency detection

The evolution of saliency detection algorithms is broadly divided into two phases. In the first phase, the algorithms address visual saliency detection by using bottom-up computation models and low-level feature representations. These methods usually make heuristic saliency assumptions, such as color contrast [5-8], boundary background [9-11] and center prior [12] to help saliency detection. Since [13], the Deep Convolutional Neural Network (CNN) based methods have drawn lots of attention due to its effective capacity for extracting rich features from raw images, and since then, the development of saliency detection entered the second phase. Compared with conventional hand-crafted low-level features, deep CNN features capture more complex high-level semantic concepts that appear in images. Based on these high-level deep features, the saliency values of the object proposals and the saliency maps can be predicted and derived effectively. In this chapter, we review some of the well-known conditional and Deep CNN-based methods, analyze their strengths and limitations, and thus make improvements.

2.1.1 Conventional semantic attention

As we have mentioned, conventional semantic attention usually takes the approach of saliency assumptions, and in the chapter, we will give a brief introduction on the two typical attention mechanisms of contrast based and boundary context based.

The contrast based method:

Contrast, which refers to the difference between color, is one of the main intrinsic image features that is widely used in image processing, and today, scientists are also using it to solve the task of SOD. The algorithm in [5] is a well-known SOD method which based on color contrast. The method determines saliency region by color contrast and outputs full resolution saliency map with well-defined boundaries of salient objects. Compared with other previous methods which can only output low resolution saliency maps, the method successfully raised saliency detection to the application level. The method sets several hypothetical constraints for the salient object detection algorithm, such as the algorithm should emphasize the largest salient objects, highlight the whole salient regions uniformly, establish well-defined boundaries of the salient objects, disregard high frequencies arising from texture,

noise and blocking artifacts, and output full resolution saliency maps efficiently. To meet the above requirements, the algorithm combine the output of several band pass filters with contiguous [ω_{lc}, ω_{hc}] pass bands, which can help to obtain a saliency map containing a wide range of frequencies. To do that, Difference of Gaussian (DoG) band pass filter is calculated as equation 2.1, then the narrow band-pass DoG filter is calculated by equation 2.2, where σ_1 and σ_2 are the standard deviations of the Gaussian ($\sigma_1 > \sigma_2$), ω_{lc} be the low frequency cut-off value and ω_{hc} be the high frequency cut-off value.

$$DoG(x,y) = \frac{1}{2\pi}\left[\frac{1}{\sigma_1^2}e^{-\frac{(x^2+y^2)}{2\sigma_1^2}} - \frac{1}{\sigma_2^2}e^{-\frac{(x^2+y^2)}{2\sigma_2^2}}\right] = G(x,y,\sigma_1) - G(x,y,\sigma_2) \qquad (2.1)$$

$$\sum_{n=0}^{N-1} G(x,y,\rho^{n+1}\sigma) - G(x,y,\rho^n\sigma) = G(x,y,\sigma\rho^N) - G(x,y,\sigma) \qquad (2.2)$$

To implement a large ratio in standard deviations, the method drives σ_1 to infinity. This results in a notch in frequency at DC while retaining all other frequencies. To remove high frequency noise and textures, gaussian filter is used. To be computational efficiency and simplicity, small Gaussian kernel $\frac{1}{16}[1,4,6,4,1]$, is used, giving $\omega_{hc} = \frac{\pi}{2.75}$. Finally, the saliency map S for an image I of width W and height H pixels can thus be formulated as:

$$S(x,y) = \| I_\mu - I_{\omega_{hc}}(x,y) \| \qquad (2.3)$$

Where I_μ is the mean image feature vector of color and luminance and $I_{\omega_{hc}}$ is the corresponding image pixel vector value in the Gaussian blurred version (using a 5×5 Gaussian kernel) of the original image, and $\|\|$ is the $L2$ norm.

The approach of computing saliency uses low level features of color and luminance, which is easy to implement and provides full resolution saliency maps. However, like most of the color contrast based salient object detection methods, its effectiveness is

limited in complex application scenarios, and in addition, the method can still only obtain the approximate boundaries of the highlighted objects.

Boundary background based method:

Many approaches try to explore the background prior to generate more accurate and robust salient maps. Boundary background is one of the commonly used as background priors. Method proposed in [9] is a representative one. Instead of considering the color contrast between the salient objects and their surrounding regions, the algorithm analyses and explores the foreground cues and background cues, then uses graph-based manifold ranking to rank the similarity of the image elements (pixels and regions) based on the cues. The graph labelling introduced by the method consist of two seemingly similar, but complementary phases. In more detail, the algorithm represents the image as a close-loop graph with superpixels as nodes, based on the affinity matrix, query the similarity of the nodes to the background and foreground, and rank them based on the similarity. In the first stage, the nodes on the four boundary sides of the image are marked as background queries, and then for each side, the saliency of other nodes is calculated based on the relevance with those background queries of the boundary side. Four side-specific maps are generated based on the four side background queries. The four side-specific maps are then integrated to generate a saliency map. In the second stage, the resulted saliency map from the first stage is binary segmented, and the labelled foreground nodes are taken as foreground queries. The saliency of each node is computed based on its relevance to foreground queries to obtain the final saliency map. Through solving the optimization problem of equation 2.4, the algorithm obtains the optimal ranking of queries. The minimum solution is computed by setting the derivative of the above function to be zero as equation. 2.5.

$$f^* = arg\min_f \frac{1}{2}\left(\sum_{i,j=1}^{n} \omega_{ij} \left\| \frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i=1}^{n} \|f_i - y_i\|^2 \right) \qquad (2.4)$$

$$f^* = (D - \alpha W)^{-1} y. \qquad (2.5)$$

Where the parameter μ balances the smoothness constraint (the first term) and the fitting constraint (the second term). The first term constraints the change between nearby points to be small. And the second term constraints the difference from the initial query assignment should not be too much.

In [9], the local smoothing cues are effectively utilized by extending the range of node connections with similarity, the geodesic distance of similar superpixels is reduced by forming a closed-loop graph, forcing the nodes on all four sides of the image to be connected so that any pair of boundary nodes is considered adjacent. The closed-loop graph effectively improves the ranking results and the performance of the algorithm.

Although the method improves the performance when significant objects appear near the image boundaries or when some background regions are not identical, it still cannot work well for images of complex scenes.

2.1.2 Deep CNN based method

The conventional SOD methods achieved good results in simple images, where simple means that there are only one or two salient objects in the image and the background of the image is relatively simple. However, in practical applications, such as in medical image analytics and industrial automation fields, the images are very complex and diverse, both the salient object and the background have countless changes, it is difficult for these methods to identify between saliency objects and backgrounds, especially when most assumptions are triggered by low-level visual cues. Thanks to the high-level semantic feature extraction ability of deep neural network, the deep learning based bottom-up saliency detection approaches have shown tremendous improvement over conventional methods.

As we known, the hand-crafted features usually represent the low-level features, which are not suitable to represent the object. While CNN can be used to obtain both low-level features and high-level features from different layers of the network. In the early days, many CNN based salient object detection methods use a combination of

local and global information to overcome the problem that local models only highlight object boundaries but not the whole objects, and global models do not preserve object details, making it difficult to detect salient objects of large size and with complex textures. In [26], L. Wang applied two subnetworks to perform both local estimation and global search (LEGS). In the local estimation phase, the model uses the original pixels as input to learn the contrast, texture, and shape information of the local images by training a neural network called DNN-L, the DNN-L responsible for assigning local significance values to each pixel considering its local context. And in the global search phase, the model takes a feature vector containing global contrast, geometric information, and the local saliency measurements evaluated by the local estimation by DNN-L as input to train another deep neural network called DNN-G to predict the saliency value of each object candidate region from a global perspective. Finally, the image's saliency map is computed by a weighted sum of the top K candidate's confidence maps. Experimental result showed that the method achieves state-of-the-art results by combining low-level saliency and high-level objectness through a supervised DNN-based learning scheme at that time. In [28], T. Wang proposed a saliency detection method that takes advantage of object-level proposals and region-based convolutional neural network (R-CNN) features. They first determine the positive and negative instances inside the image and measure the confidence by R-CNN, then train a feature projection matrix to project high-dimensional features into a low-dimensional subspace to remove the redundant information in the high-level feature space.

With the development of the backbone of the deep neural network for object detection, such as the VGG and Resnet, and the development of the feature fusion methods, the end-to-end Deep CNN (DCNN) model based approaches are proposed for salient object detection. Guanbin Li and Yizhou Yu [50] proposed an end-to-end deep contrast network for salient object detection. The network includes two sub-network, one sub-network is multi-scale FCN based for pixel level saliency map generation, it can generate a one-eighth resolution saliency map of the original image by exploiting the visual contrast of multi-scale convolutional layers. The sub-network is designed with deep enough layers to produce multi-level features, so that to enable

the detection of remarkable objects at different scales and find subtle visual contrasts inside the image. Considering the salient objects have irregular shapes, the other sub-network is designed to generate the super pixel saliency map to enable the saliency map can represent the visual contrast of area and area boundary. Then the two maps are fused to get the final saliency map. N. Liu and J. Han in [51] proposed a deep hierarchical network for salient object detection called DHSNet. The network hierarchically detects objects from global view to local content, refines from coarse to fine dimensions, explores all potential saliency cues, and efficiently preserves object details. The network consists of two parts, the GV-CNN and HRCNN, where GV-CNN contains 13 convolutional layers, 4 pooling layers, a fully connected layer, and a reshaping layer for learning various global structural saliency cues, and HRCNN consists of 4 recursive convolutional layers and 4 down sampling layers for fusing local contextual content and adjusting the details of the saliency map in a hierarchical and progressive manner. Overall DHSNet [51] adopts a hierarchical and progressive strategy of optimal integration to learn from global perspective to local content.

In recent years, the research on deep learning model design with attention for saliency object detection are receiving more and more attention. [52] proposed a pyramid feature attention network (PFAN) to enhance the high-level context features and the low-level spatial structural features. The method first uses a context-aware pyramid feature extraction (CPFE) module to extract multi-scale low level and high-level features. Then introduce a channel-wise attention (CA) model in high-level feature extraction layers and a spatial attention (SA) model in low-level feature extraction layers to further enhance the effectiveness of the feature extraction. Finally, the output from CA and SA are fused for saliency map generation. Xu, B. etc have proposed another attention architecture in [53] for saliency object detection. The architecture contains three parts, the coarse locating module (CLM), attention-based sampler and a fine segmenting module (FSM). CLM is responsible to locate rough areas containing salient objects without boundary detail, the attention-based sampler highlights the salient object regions and generate a body-attention label. Then the FSM is responsible to obtain the precise segmentation of the salient object with the original image and the body-attention label as input.

2.1.3 The development and requirement in industry

Currently, algorithms for object detection and recognition, image and video compression, video summarization, photo stitching/media repositioning, cropping and stapling, image quality assessment, image segmentation, content-based image retrieval and image acquisition browsing, image editing and processing, visual tracking, object discovery and human-computer interaction have been widely adopted in the technological fields of computer vision, graphics and robotics. These technologies are adopted in a wide range of applications such as retail, security, road surveillance, manufacturing, and healthcare. As the foundational pre-processing module of these technologies, salient object detection techniques are critical. In particular, the importance of saliency detection is becoming increasingly evident with the rapid development of 5G and IoT technologies.

In recent years, AI-assisted medical image detection and diagnosis is rapidly developing, especially in the fields of radiology and pathology, where AI plays an important role. However, there are many difficulties in this process. Here we take digital pathology image analysis as an example. The first difficulty is related to the computational bottleneck of image processing. The task of digital pathology medical image analysis is data-intensive and computationally intensive, with high-resolution whole-slide images including thousands of tile images and hundreds of whole-slide images. All of them need to be analysed before reaching a final conclusion, and even if parallelization has been implemented to speed up the computational process, it only partially solves the speed problem. The second difficulty is due to the lack of a powerful and efficient algorithm for processing large-scale image analysis. Since an whole slide image consists of thousands of tile images, which can easily generate millions of objects with tens of millions of feature values, it is a very painful process for physicians to manually browse through these tile images for diagnosis. An experienced physician usually first perform a global analysis on large scale of the images to identify areas of interest and then perform a local focus analysis. Obviously, we need a visual attention mechanism in the algorithm design to automatically find and focus on regions of interest during algorithm execution, especially for large-scale image batch data, such

as whole slide images, where fast discovery of regions of interest is critical. In addition to the difficulties mentioned above, there are other minor issues that result from dealing with large numbers of images during image processing and analysis development. In summary, although computer-aided diagnostic applications should theoretically benefit from the growing size of medical image data, the development of image processing and analysis development tools is technically a rather challenging task.

We focus our AI application field on data analysis of medical images for the following reasons. First, medical images with metadata and diagnostic reports can be easily de-identified and ready for open access without privacy concerns. In fact, there are more and more open access medical image databases available on internet to facilitate computer aided diagnosis researches. Second, medical image data dominate all types of medical data in size with over 80%. Effective AI-assisted diagnosis and analysis will greatly reduce the rate of missed diagnosis and misdiagnosis by doctors and reduce their burden, while medical institutions can generate commercial value from it and compensate for the high cost of long-term archiving due to regulatory requirements. Third, medical image data formats are relatively fixed, with limited variation in DICOM standards compared to other clinical data types, thus requiring less engineering effort to handle data preparation for ingestion into analytics platforms. In conclusion, edical image big data analytics is with higher business demand and lower technical barrier thus becomes our obvious research focus.

2.2 The proposed semantic attention for saliency detection

2.2.1 Research motivation

Visual attention is an important mechanism in the human visual system. It is observed that it not only allows the brain to observe objects under the guidance of low-level image attributes, but also to describe images semantically at a high-level. In particular, people do not describe everything in an image; they tend to talk more about semantically meaningful regions and objects in the image. This observation inspired researchers to design visual attention algorithms that can focus not only on low-level features, but also on high-level semantic features, just like humans.

Object proposal technology was first introduced to visual saliency detection by Chang et al. [1] and boosted by [2-3]. Object proposal technology generates a set of class-independent image regions and assigns each region a score to indicate the probability of a region being an object, the generated regions are called proposals. The proposals can preserve features of object-level knowledge and boundary cues by segmenting all the objects or some of the objects from the input image, and these features can be used for saliency detection. However, the object proposal techniques usually generate a large number of cropped or background regions along with the salient object proposals, it is hard to use the proposals directly for training the downstream models for saliency detection. In addition, the original probability scores cannot be directly utilized to measure the saliency values of regions. As shown in Fig 2.1, the object proposals with red rectangular bounding boxes are fragments of the object "bird" and "coach", both of them have high probability scores, but these proposals cannot be regarded as salient regions and the probability scores cannot be used because they belong to the background of the image.

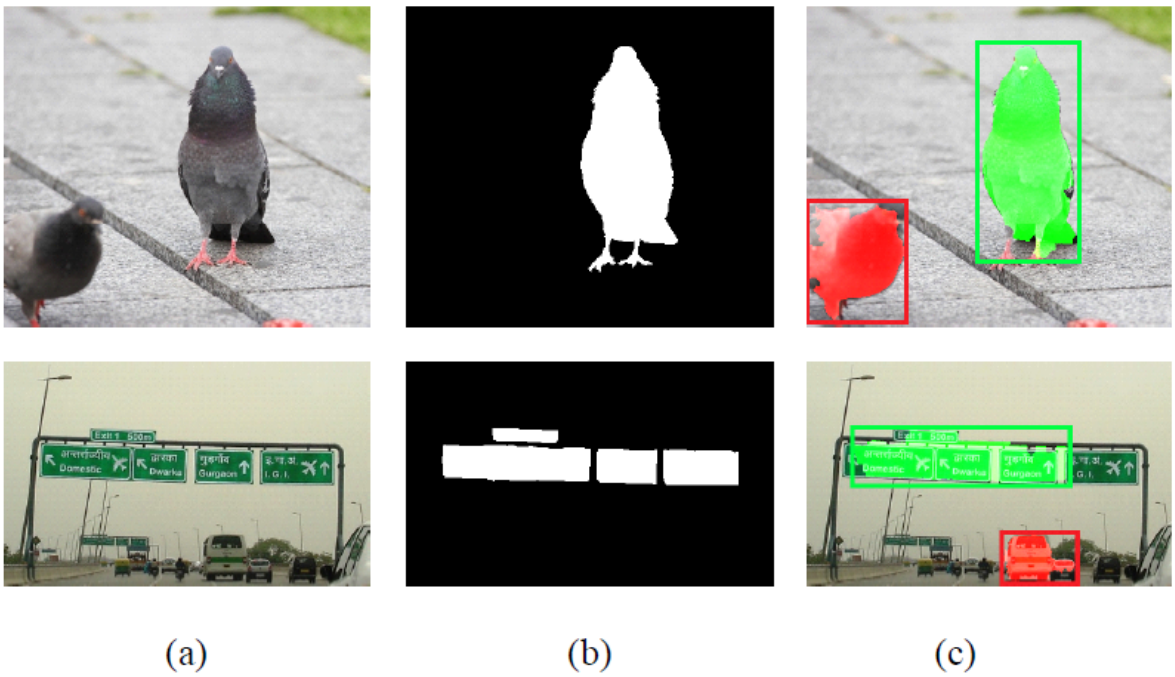


Fig. 2.1. (a) Input image. (b) Ground truth saliency map. (c) Object proposals generated by [4]. The red proposals have large probabilities to be an object, but they are less salient regarding to the input images.

To address this problem and to leverage object proposals in visual saliency detection, we designed an effective proposal-aware semantic attention method to solve this problem. First, a proposal selection module is designed to select meaningful proposals and remove unnecessary ones. Then a re-weighting mechanism is designed to obtain a reasonable saliency score for each proposal.

2.2.2 Semantic attention mechanism

It is observed that regions with high attentions in an image usually present a similar semantic concept to that of the salient objects. Based on this observation, we can implement saliency detection with an object detection model that assigns attention scores to each object region in an image. Faster R-CNN [14] is recognized as a classical object detection algorithm, which is a proposal-based two-stage object detection, and it can predict a large number of objects with objective scores. Therefore, in our solution, we leverage the Faster R-CNN model for the saliency detection task by training the model to estimate the saliency scores of each bounding boxes. Since the bounding boxes only provide rough locations of objects, we applied a fine-grained segmentation module on the bounding boxes to separate the salient objects out. The deep features of masked objects and object proposals are extracted from the fully connected layer. Instead of generating the saliency maps based on the masked objects, which can cause the loss of some subtle object parts during the segmentation, the attentions of object proposals are calculated according to their feature distances and overlaps with the masked object and for the lost object parts, they are constructed by conjoining multiple object proposals. By incorporating more human semantic knowledge on the training process, the deep features are capable of representing objects and generating more accurate attention values. Salient object proposals are selected according to their semantic attention probabilities. And the final saliency map is calculated by the weighted average of selected object proposals.

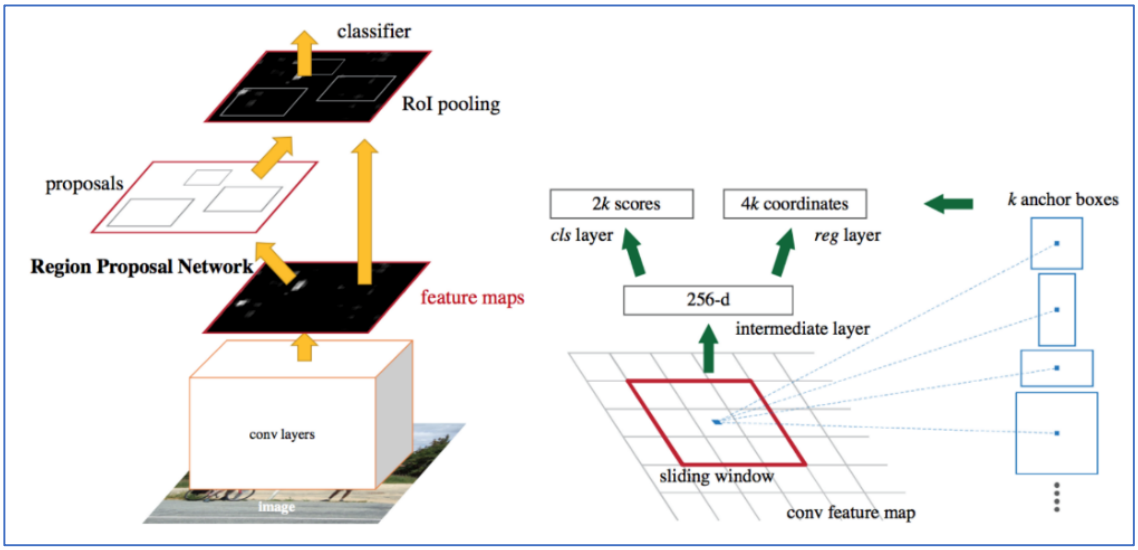


Fig. 2.2 An illustration of Faster R-CNN. (Image source:[14])

Figure 2.3 shows the framework of the algorithm. The processing can be divided into three steps. First, select M candidate proposals, then perform proposal selection, after that reweighting the proposals and obtain the final saliency map. The geodesic object proposal (GOP) [8] algorithm, due to its fast processing and high accuracy, is used for candidate proposal generation. The generated proposals are ranked according to the reasonableness of the region size, the top M proposals (about 70%) are selected, and the rest are eliminated. The selected candidates are processed in two different ways. On one hand, these candidates are directly segmented by GrabCut [2] to obtain M salient object mask map. On the other hand, these candidates are input into a binary classification model Faster-RCNN [1] to obtain the object proposals (N bounding boxes), then for each object proposals perform GrabCut [2] to derive object salient mask map. All the salient mask maps and the object salient mask maps will be input into another Faster-RCNN [1] model to extract the deep features with high-level semantics. These deep features will be used to generate the probability map.

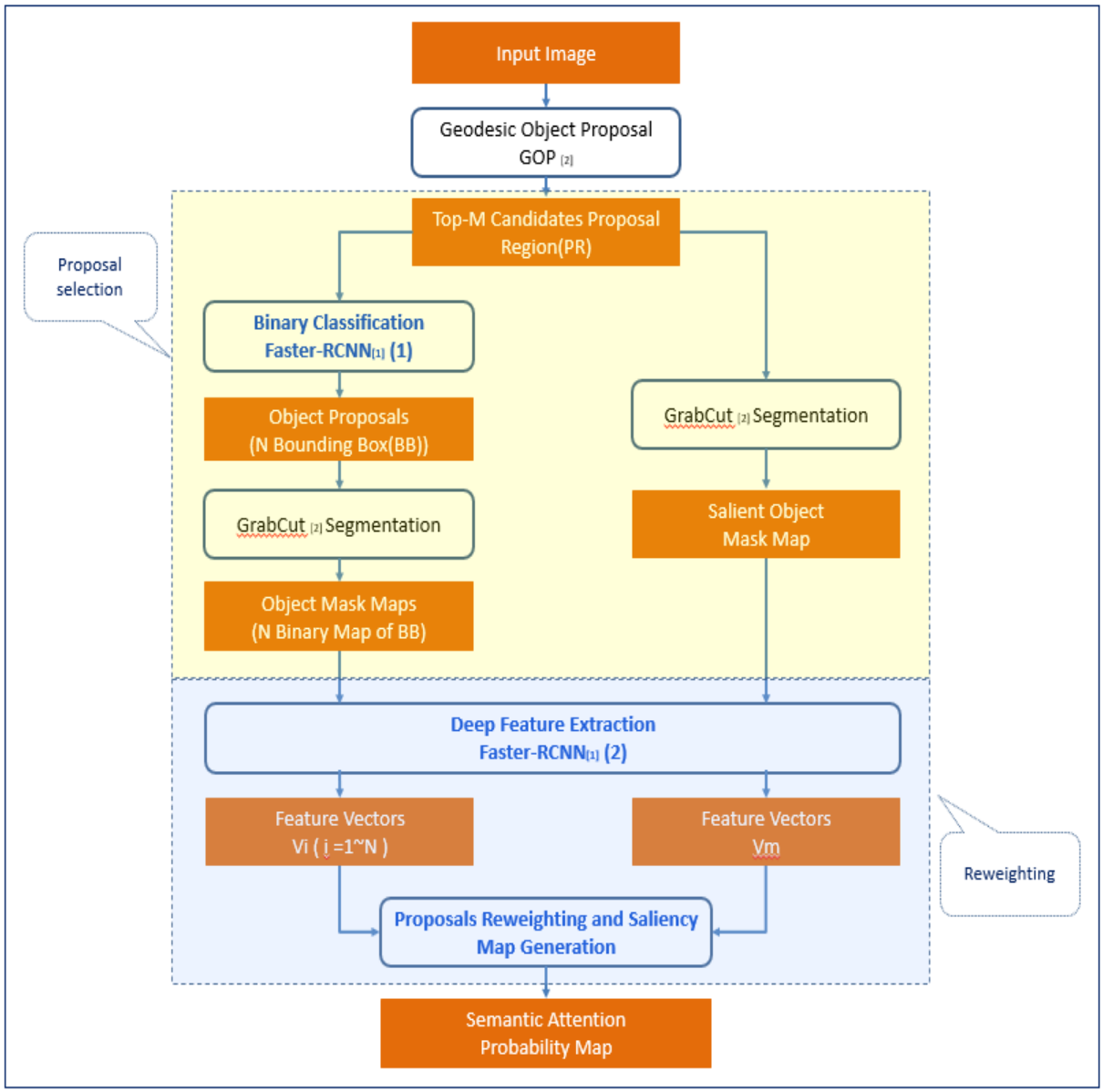


Fig. 2.3. Algorithm Framework

2.2.2.1 Proposal selection

Generate Object Proposals

The geodesic object proposal (GOP) method [8] takes an image and its boundary probability map as input and then identifies a set of foreground seeds to generate foreground and background masks. Each pair of masks specify an object proposal and generate a score to indicate its probability of being an object. Given an input image we use the GOP to generate a set of object proposals.

One problem is that some proposals generated by GOP are either too small which cannot be valid segments or too large with strange aspect ratios. We refine the proposal candidate set by filter out unreasonable cases. Proposals whose overlap with the whole image under 2% or bigger than 70% or their own aspect ratio less than 10% are not counted.

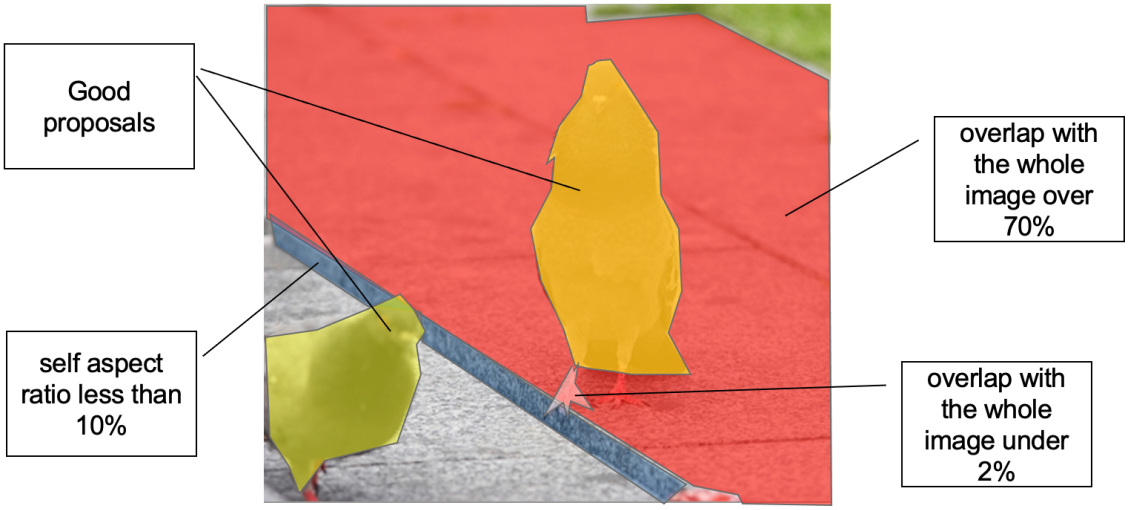


Fig. 2.4. Example of unreasonable proposal candidates

After refinement, we select top N candidates with the largest probabilities for computational efficiency and denote them as $Ri, i \in \{1 \dots N\}$. The selected object proposals cannot be directly utilized to construct saliency maps.

In existing saliency detection datasets, it is common that objects locate on the image corner and only show a small tip, some subtle objects even scatter around the image. All these objects cannot be regarded as salient. For proposals centering on these kinds of objects, their original probabilities of being an object are not suitable to measure the saliency. In order to address these kind issues, we propose a semantic attention model to re-weight the object candidates.

Semantic Attention Model

The proposed semantic attention model which calculates proposals' attention probabilities by comparing their feature distances with the possible salient object.

First, we estimate the location and shape of the possible salient objects by the Grab-cut segmentation method, and thus obtain a set of object proposals.

Then, we fine-tune the Faster R-CNN network to support the salient object detection task. The features extracted from the Faster R-CNN can capture high-level semantic visual patterns and are more robust to complex backgrounds.

We extract the deep features by Faster R-CNN model from both the object proposals and the possible salient object for saliency prediction. Since object proposals that are close to salient objects in the feature space should be paid more attention, the semantic attention of each object proposal is predicted according to its feature distance and overlap with the possible salient object.

To take advantage of the deep feature extraction capability of the Faster R-CNN, while giving the model the ability to detect salient objects rather than just object detection, the following improvements were made to the model.

- Replace the last layer of Faster RCNN by a FC layer and a sigmoid function for visual saliency detection.
- Use Region Proposal Network (RPN) sub-network to generate Bounding Boxes, acquire the foreground Bounding Boxes with intersection over union (IoU) > 0.7, and background Bounding Boxes with IoU < 0.2 as the generated training data. Figure 2.5 shows an example of the generated training data.
- Adopt the binary cross-entropy loss function to train the model.

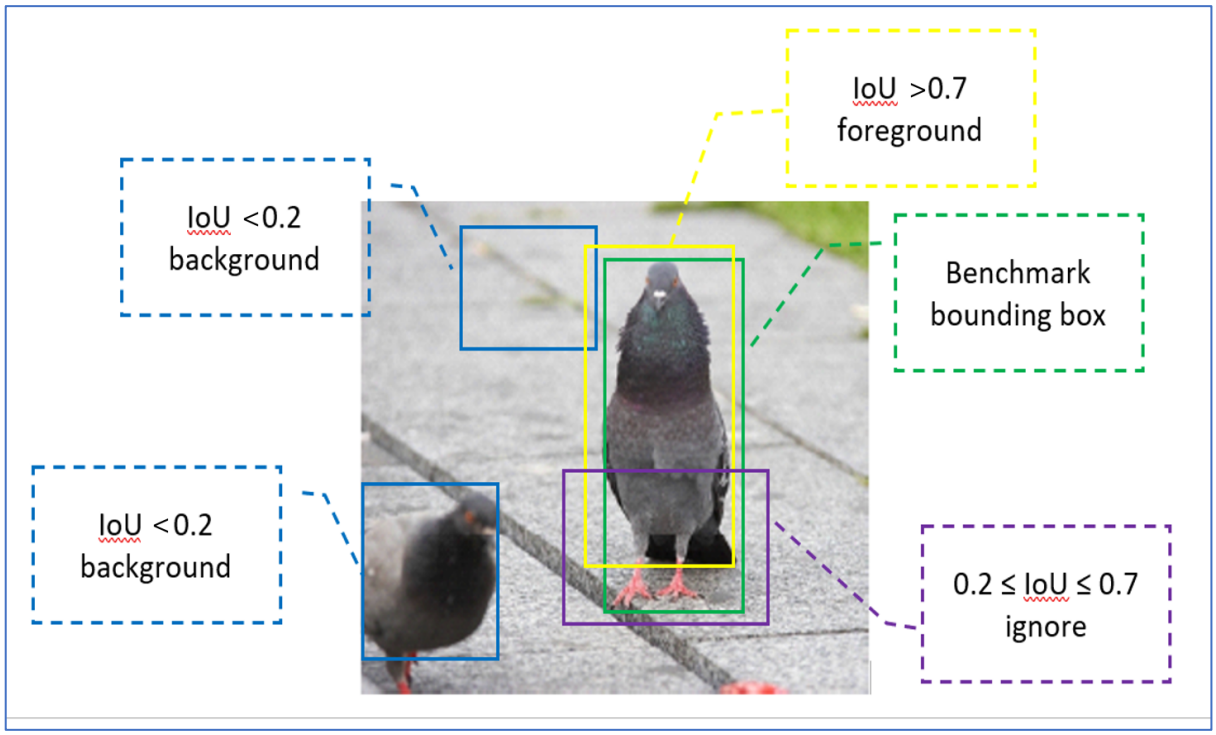


Fig 2. 5 Example of training data generation

Training Faster R-CNN.

Faster R-CNN network [14] is proposed for object detection. It takes an image as input and outputs a set of bounding boxes and their probability estimates over different object classes. Due to the efficiency and effectiveness of Faster R-CNN, it is applied to visual saliency detection in our study. Different from the task of object detection to detect and discriminate objects of various categories, saliency detection aims at separating general foreground objects of all classes. In this case, a group of rectangular boxes are generated through the RPN network. The rectangular boxes are assigned class labels by categorizing them into three classes according to the intersection over union (IoU) value with salient objects. For more detail, each box will be assigned a class label y. y = 1 means current rectangle is positive and its IoU with the salient objects is larger than 0.7. y = 0 represents negative samples with the IoUs smaller than 0.2. Rectangles with 0.2 < IoUs < 0.7 are excluded from the training set to avoid the neural network being confused. Examples are showed in Figure 2.5. where benchmark bounding box represent the salient object and the IoU is derived based on the bounding box.

The saliency detection is formulated as a binary classification problem. We take the pre-trained VGG16 model [15] as backbone network and fine-tune it on the saliency detection task. The last layer of Faster R-CNN is replaced by a fully connected layer and a sigmoid function. The output is a scalar between 0 and 1 to represent the salient degree. We train neural network by minimizing the following binary cross-entropy loss defined in equation 2.6,

$$L = -\frac{1}{N_B}\sum_{i=1}^{N_B}[y_i log s_i + (1 - y_i)\log(1 - s_i)] \qquad (2.6)$$

where y_i and s_i are the target label and predicted saliency score of B_i. N_B is the number of selected training samples.

2.2.2.2 Proposal reweighting

Masked Object.

Faster R-CNN can predict the attractive values of object rectangles and k most salient ones are fused to estimate the location of objects. However, coarse bounding box locations do not contribute to fine-grained saliency detection, most of the detailed shape and contour information are missing. Additional, feature representations generated by Faster R-CNN network contains plenty of background noises. As shown in the third column of Fig 2.6, backgrounds occupy most area of the bounding box and objects features generated in these cases might be misguided by noisy backgrounds. To obtain objects with accurate shape and contour, one simple strategy is to segment the regions out from bounding boxes. This is realized by a GrabCut method [16], which generates binary maps to indicate the object and background regions. Some examples of masked objects are shown in Figure 2. 6 (d). Compared with separating objects from the whole image directly, we first estimate the possible location of foregrounds and then perform a fine-grained segmentation. The position limitation of bounding boxes avoids GrabCut being disturbed by irrelevant image areas and improves the accuracy of salient object detection. The mask maps are further utilized to calculate semantic attentions.

Retrain Faster-RCNN model by using object segmentation mask on the training dataset can remove the background noise and obtain objects with accurate shape and contour.

Semantic Attention Probability.

To evaluate the attractive probabilities of regions, we measure these regions' semantic similarities with the masked objects. Although the mask maps can indicate possible locations of salient objects, there are still some subtle parts are misclassified. Take the masked images in Figure 2.6 as an example, in the first case, the background grassland between sheep's two legs is mis-regarded as foreground. In addition, when objects have similar appearance with backgrounds, it is more challenging to extract the foreground objects. The binary prediction deteriorates this situation by assigning absolute confidences, either 1 or 0, to misclassified regions. To address the problem, instead of directly utilizing the masked map, we extract the deep features through Faster R-CNN model to predict the attention probabilities of object proposals $Ri, i \in \{1 \dots N\}$ which generated in Section 2.1. Since the object proposals may contain all or part of the shape and contour information of the objects, fusing them together can avoid the generated saliency map being affected by misclassified regions in the masked maps.

Given a rectangular boxes Bi generated by Faster R-CNN, we segment out possible objects according to its GrabCut result and replace the background regions by the mean of B_i. The new generated rectangular is termed as B'_i. To extract more discriminative feature representations of regional proposals, the Faster R-CNN is fine-tuned with B'_i as training samples. B_i and B'_i have the same class labels. During the test phrase, we perform the same operation on the mask map R_M and extract its feature vector v_M through the fine-tuned Faster R-CNN. Similarly, the features vectors of N object proposals are denoted as $v_i, i \in \{1 \dots N\}$. Different from low-level contrast feature, deep CNN features capture high-level semantic concepts appearing in the images and are more powerful in representing images. Proposals that have similar semantic understanding with the masked object are more potential to be salient, so

the semantic attention should highlights the proposals that are similar with the masked object in understanding high-level visual patterns. To generate the final saliency map, we formulate the semantic attention of each candidate proposal as follows:

1. Calculate the semantic attention of each candidate proposal as equation 2.7,

$$a_i' = \frac{\exp\left(-\left||v_i - v_m|\right|_2^2\right)}{\sum_{j=1}^{N} \exp\left(-\left||v_j - v_m|\right|_2^2\right)}, \; for\; i = 1, \; \ldots, N \tag{2.7}$$

   where ‖:‖ denotes the norm-2 distance of two vectors.
2. Sort the candidate proposals according to their attention score a_i' in a descend order;
3. Select top-N′ proposals to formulate the saliency map via weighted summation.

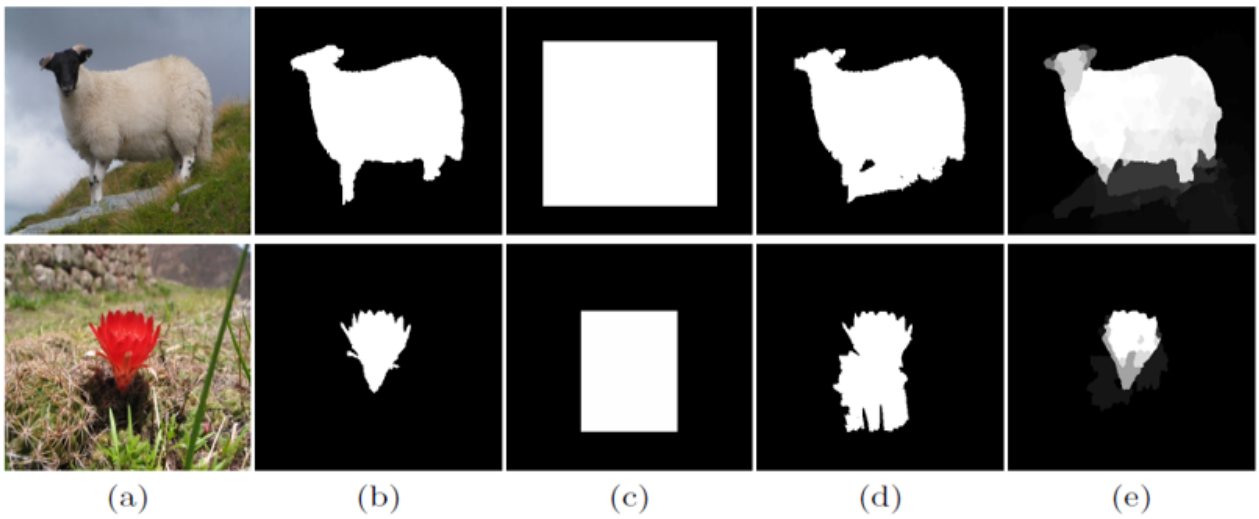


Fig. 2.6 Example maps generated in the semantics attention module. (a) Test image, (b) ground truth, (c) bounding box generated by Faster RCNN, (d) masked object, (e) semantics attention map.

2.2.3 Simulation

Evaluation.

In addition to directly comparing the saliency map with ground truth, we evaluate the performance using the precision-recall (PR) curves, F-measure and Area Under Curve (AUC).

Precision-recall curve calculation:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Where TP (true positive) means that a test result that correctly indicates the presence of a condition or characteristic; TN (true negative) means that a test result that correctly indicates the absence of a condition or characteristic; FP (false positive) means a test result which wrongly indicates that a particular condition or attribute is present, and FN (false negative) means a test result which wrongly indicates that a particular condition or attribute is absent.

Given a saliency map, the precision and recall are computed by segmenting the map with a threshold to construct a binary map and comparing it with the corresponding ground truth. Taking the average of precision/recall pairs of all images, we plot the precision-recall curve at different thresholds.

F-measure is an overall performance indicator computed by the weighted harmonic of precision and recall,

$$F_r = \frac{(1+r^2) \times Precision \times Recall}{r^2 \times Precision + Recall},$$

where r is set to 0.3 to weigh more on precision as suggested in [5].

To better assess the results, we also introduce the AUC which calculates the area under the ROC curve. The ROC curve uses false positive rate (FPR) as the X-axis and true positive rate (TPR) as the Y-axis, each point represents the different FPR and TPR obtained by setting different threshold values, and finally plotted as a curve. When using normalized units, AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Therefore, the higher the AUC, the higher the accuracy of the classifier will be.

$$Specificity = \frac{TN}{TN + FP}$$

$$FPR = 1 - Specificity = \frac{FP}{FP + TN}$$

$$TPR = Sensitivity = \frac{TP}{TP + FN}$$

Dataset.

Four benchmark datasets are selected for method evaluation.

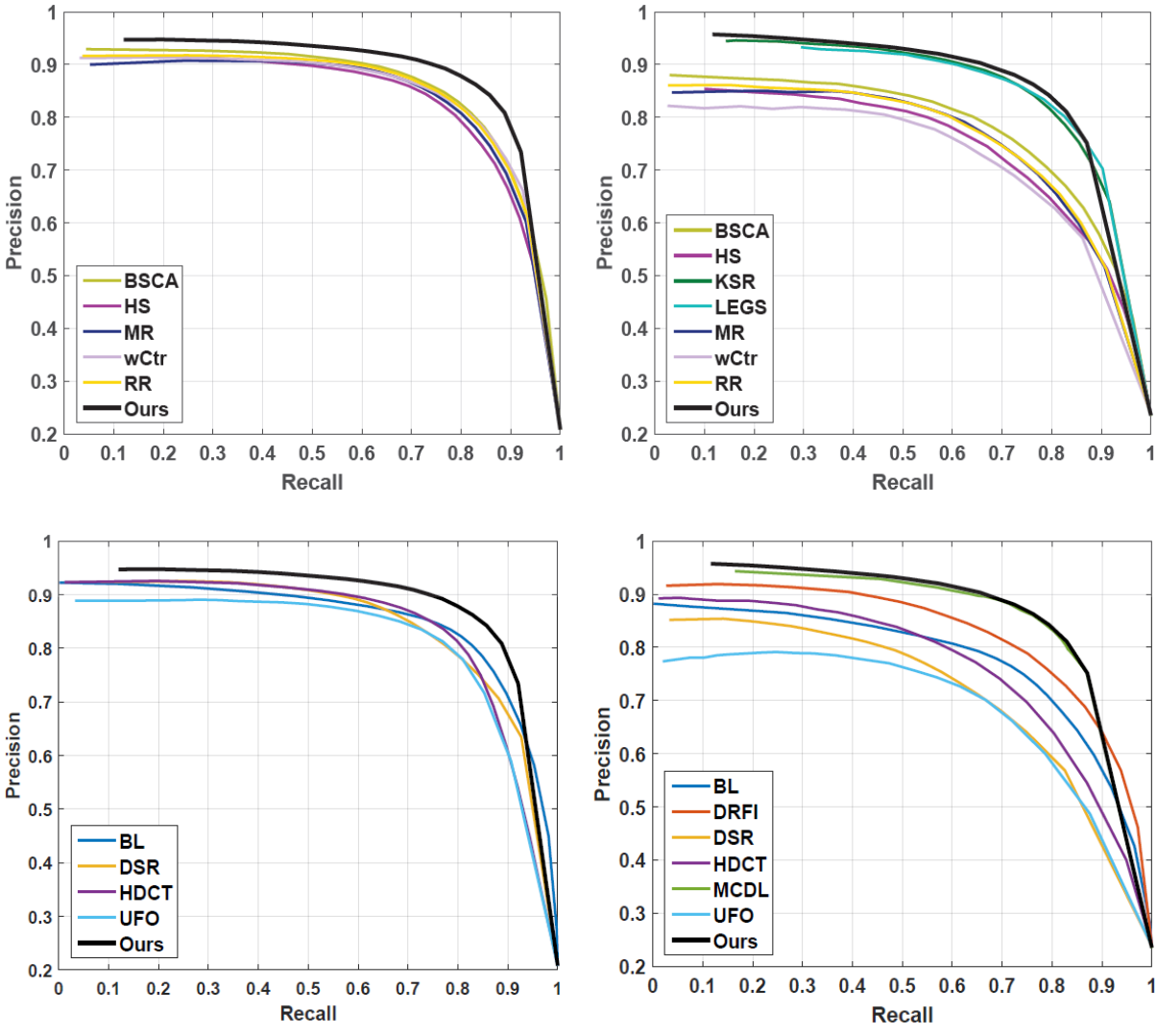
- MSRA-5000 [17], the dataset has 5000 images with saliency annotations, which is widely used for saliency detection.
- ECSSD [8], the dataset proposed by Yan et al. [8] has 1,000 images with various complex scenes from the Internet.
- SOD [18], the dataset consists of 300 challenging images with multiple objects of various sizes and locations, and
- PASCAL-S [19] contains 850 images with cluttered backgrounds and multiple objects.

Implementation details.

1. The GOP method [4] generally produces 1,000 object proposals for each image.
2. We filter out unreasonable ones and select the top N = 200 proposals according to their GOP scores.
3. In the semantic attention model, we fine-tune Faster R-CNN on the DUT-OMRON dataset [10]. The initial learning rate is 0.001 and decreases by a factor of 0.1 every 10k iterations. The momentum parameter and weight decay are 0.9 and 0.0005 respectively. The training process converges after 200k iterations with the stochastic gradient descent optimizer.
4. The top N′ = 15 candidates are fused to generate the final saliency map.

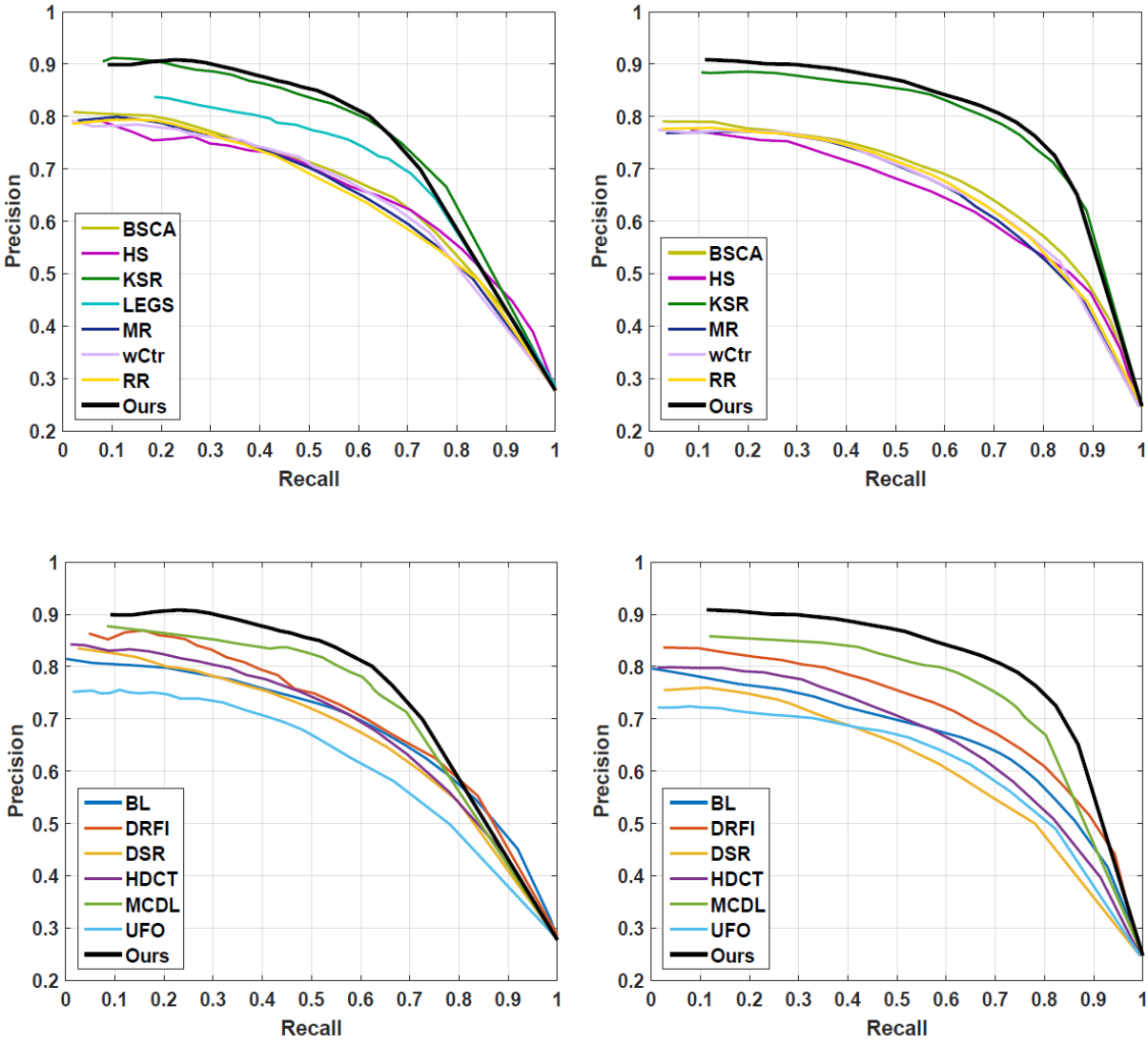
Comparison.

Semantic attention model can locate salient regions with great accuracy and highlight the whole object regions with unambiguous boundaries. And the algorithm outperforms 13 saliency detection methods.



(a) MSRA-5000

(b) ECSSD



(c) SOD (b) PASCAL

Fig. 2.7. Quantitative comparisons of proposed approach and 13 baseline methods on four saliency detection datasets.

*	MSRA-5000		ECSSD		SOD		PASCAL-S	
	F-measure	AUC	F-measure	AUC	F-measure	AUC	F-measure	AUC
KSR	-	-	0.7817	0.9251	0.6703	0.8510	0.7039	0.8970
LEGS	-	-	0.7887	0.9239	0.6492	0.8117	-	-
MCDL	-	-	0.7959	0.9186	0.6789	0.8176	0.6912	0.8699
BL	0.7838	0.9360	0.6825	0.9147	0.5723	0.8503	0.5668	0.8633
BSCA	0.7933	0.9428	0.7046	0.9176	0.5852	0.8364	0.6006	0.8665
RR	0.8072	0.9089	0.6577	0.8283	0.5665	0.7888	0.5873	0.8251
HDCT	0.7733	0.9318	0.6897	0.9039	0.6108	0.8504	0.5824	0.8582
wCtr	0.7960	0.9169	0.6774	0.8779	0.5978	0.8014	0.5972	0.8433
DRFI	-	-	0.7337	**0.9391**	0.6031	0.8464	0.6159	0.8913
MR	0.7575	0.9044	0.6932	0.8820	0.5697	0.7899	0.5881	0.8205
DSR	0.7760	0.9247	0.6636	0.8604	0.5968	0.8210	0.5513	0.8079
HS	0.7668	0.9043	0.6363	0.8821	0.5210	0.8145	0.5278	0.8330
UFO	0.8011	0.8950	0.6442	0.8587	0.5480	0.7840	0.5502	0.8088
Ours	**0.8407**	**0.9540**	**0.8012**	0.9388	**0.6824**	**0.8514**	**0.7213**	**0.9131**

Tab. 2.1. Quantitative comparisons in terms of F-measure and AUC. The best results are annotated in overstriking fonts.

We compare the proposed saliency detection model with 13 state-of-the-art methods, including MCDL [20], BL [21], BSCA [22], HS [8], UFO [3], wCtr [11], DRFI [23], DSR [24], HDCT [25], LEGS [26], MR [10], RR [27], KSR [28]. For fair comparison, the saliency maps of different methods are either provided by the authors or achieved by running available codes or software.

We perform comparisons of the proposed algorithm and the state-of-the-art saliency detection methods on four datasets as shown in Table 2.1 and Figure 2.6. The F-measure and AUC shown in Table 2.1 demonstrate our approach achieves the best results under almost all measurements. The KSR [28], LEGS [26], MCDL [20], and DRFI [23] baselines randomly select images and ground truth from MSRA-5000 dataset for training and test, we do not report the experimental results of these four algorithms on the MSRA-5000 dataset for fair comparisons. In addition, the LEGS [26] takes about 300 images from the PASCAL-S dataset as training samples, its results on the PASCAL-S dataset are not included. In Figure 2.7, each column represents the precision-recall curves on one dataset. The comparison results in terms of PR curves verify that our proposed method achieve state-of-the-art performance on four datasets. We present some example saliency maps generated by the proposed method and ten

state-of-the-art algorithms in Figure 2.8. Our method can locate salient regions with great accuracy and highlight the whole object regions with unambiguous boundaries.

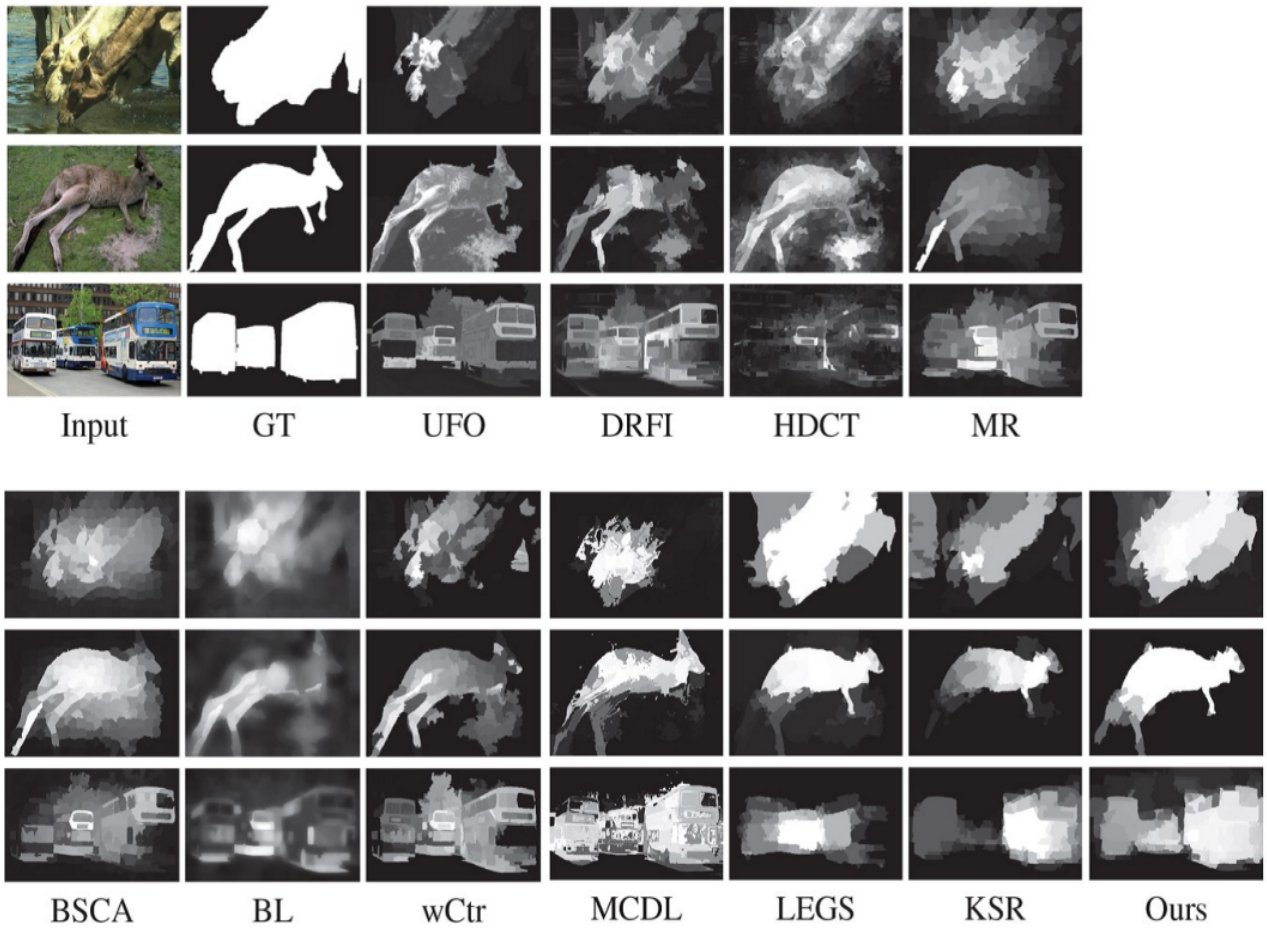


Fig. 2.8. Example saliency maps generated by proposed approach and some state-of-the-art methods

2.2.4 Conclusion

The proposed semantic attention model calculates proposals' attention probabilities by comparing their feature distances with the possible salient object. It estimates the location and shape of salient objects by the Grab-cut segmentation method. Fine-tune the Faster R-CNN network on the saliency detection task to extract deep features. These features capture high-level semantic visual patterns and are more robust to complex backgrounds. Since object proposals that are close to salient objects in the feature space should be paid more attention, the semantic attention of each object proposal is predicted according to its feature distance and overlap with the possible salient object.

The major contributions of our work are as follows:

- A new framework for visual saliency detection where salient proposals are dynamically separated out according to semantic attention probabilities.
- The semantic attention mechanism is robust to images with complex background due to the consideration of high-level semantic concept.
- The attention mechanism combined the bottom-up attention and top-down attention mechanism.
- The algorithm outperforms 13 saliency detection methods.

2.3 The proposed visual attention in Deep Learning

2.3.1 Research motivation

The attention mechanism in deep learning is designed to draw on the human way of visual attention. Human usually scan the entire image quickly, then focus on selective areas with salient objects and capture detailed information in these regions for the following analytics and processing.

On one hand, the attention mechanism in DL is essentially the same as the human selective visual attention mechanism. The main goal is to automatically select information that is more critical to the current task and suppress irrelevant information through model design. In this way, more detailed information of salient objects can be obtained, and the efficiency and accuracy of visual information processing tasks are improved. On the other hand, computing power is still a bottleneck restricting the development of neural networks today. Although optimization operations such as local connection, weight sharing, and pooling can make the neural network simpler and effectively alleviate the contradiction between model complexity and expressive ability, the amount of information to be memorized is still high. By introducing attention mechanism in deep learning, the amount of information processed can be reduced, thereby reducing the required computing resources and the information that needs to be memorized.

2.3.2 Overview of attention mechanism in Deep Learning

In recent years, most of the research work on combining deep learning with visual attention mechanisms has focused on the use of masks to form attention mechanisms. The principle of masking is to identify the key regions or features in an image by introducing a layer of weights to the network. Through training, the deep neural network learns to identify the areas of the picture that need attention, also known as the attention map.

Attention mechanism in deep learning can be divided into soft attention, hard attention and self-attention. The key point of soft attention is that this kind of attention is more focused on regions or channels, and it is deterministic attention, after learning, the attention map can be generated directly through the network. The most critical point is that soft attention is differentiable, and it can be learned by neural networks to calculate the gradient and get the attention weight by forward propagation and backward feedback. Hard attention is different from soft attention in that hard attention is a stochastic prediction process with more emphasis on dynamic changes, and it is a non-differentiable attention, and the training process is often done by reinforcement learning (reinforce learning).

In computer vision, the attention mechanism is usually divided into spatial domain, channel domain, and mixed domain in terms of attention domain. These domain based attention methods have been developed in deep neural network for the tasks of saliency object detection and semantic segmentation.

Spatial attention module.

The spatial attention module was first introduced to CNN network as “Spatial Transformer Network” by Max Jaderberg et al, in 2015 [29]. As we know, the CNN network is not invariant to large-scale spatial transformations, which will greatly limit the performance of the CNN network. The introducing of spatial attention module is to enable the translation invariance, rotation invariance and scaling invariance on the existing convolutional network to improve the performance. The spatial attention module uses the inter-spatial relationship of the features to generate a spatial attention map, then guides the CNN model to pay attention to the meaningful feature spatial

position through it. Figure 2.9 shows the architecture of spatial transformer network module. As introduced by the author, the input feature map U is passed to a localisation network which regresses the transformation parameters θ. The regular spatial grid G over V is transformed to the sampling grid $T\theta\ (G)$, which is applied to U, producing the warped output feature map V . The combination of the localisation network and sampling mechanism defines a spatial transformer.

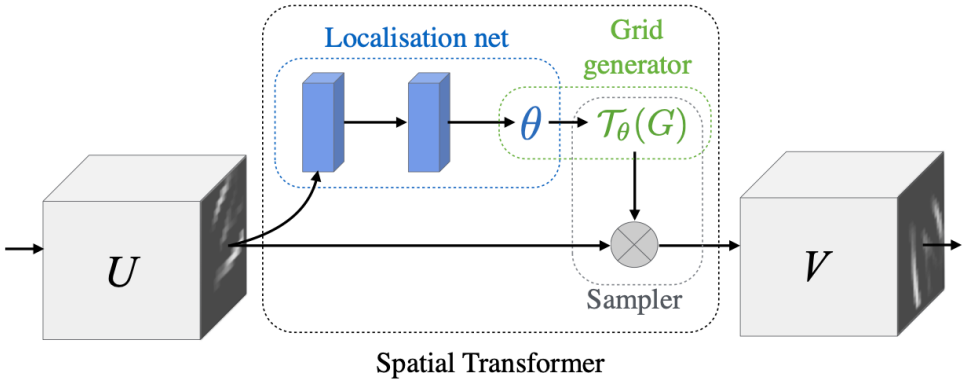


Fig. 2.9. Architecture of a spatial transformer module (Image source: [29])

Another well-known spatial attention module is introduced by Woo et al. in CBAM: Convolutional Block Attention Module [30]. The spatial attention module has been shown in figure 2.10. To compute the spatial attention, they first apply average-pooling and max-pooling operations along the channel axis and concatenate them to generate an efficient feature descriptor. On the concatenated feature descriptor, they apply a convolution layer to generate a spatial attention map $M_s(F) \in R^{H\times W}$ which encodes where to emphasize or suppress. Then they aggregate channel information of a feature map by using two pooling operations, generating two 2D maps: $F^s_{avg} \in R^{1\times H\times W}$ and $F^s_{max} \in R^{1\times H\times W}$. Each denotes average-pooled features and max-pooled features across the channel. Those are then concatenated and convolved by a standard convolution layer, producing the 2D spatial attention map. In short, the spatial attention is computed as:

$$M_s(F) = \sigma(f^{7\times 7}([AvgPool(F); MaxPool(F)]))$$

$$M_s(F) = \sigma(f^{7\times7}[F^s_{avg}; F^s_{max}]))$$

where σ denotes the sigmoid function and $f^{7\times7}$ represents a convolution operation with the filter size of 7 × 7.

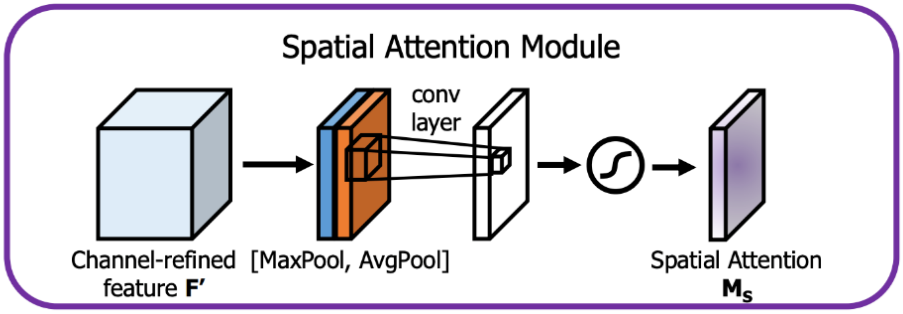


Fig. 2.10. Spatial Attention Module (Image source: [30])

Channel attention module.

The channel attention module is designed for channel-based attention in convolutional neural networks. The feature map of each channel has different attributes. For example, some channels mainly contain object features, and some contain background features. Through channel attention module, assign different weights to different channel according to its feature attributes. The channel attention map will be generated by exploiting the inter-channel relationship of features. Squeeze and Excitation (SE) Module [31] and Channel Attention Module (CAM) [30] are two typical channel attention module design.

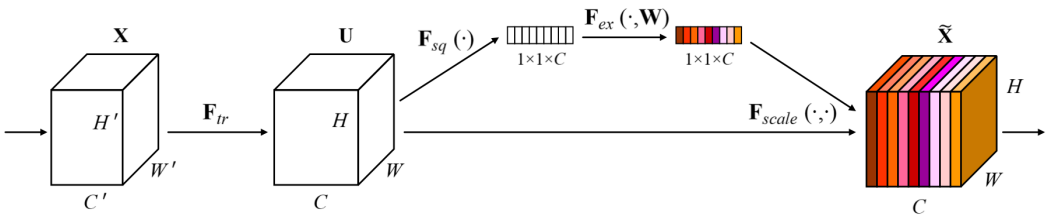


Fig. 2.11 A Squeeze-and-Excitation block. (Image source: [31])

As shown in Figure 2.11, the SE module [31] includes two stages: the Squeeze and Excitation. In brief, the squeeze part is realized through Global Average Pooling, which

is to average all the spatial information in a channel into one value. The Excitation part is composed of 2 full connections, the first full connection realizes channel compression and reduction calculation, and the second full connection restores the number of channels. Full connection is used to discover the correlation between channels to obtain a real scale.

The CAM [30] module shown in Figure 2.12 aggregates the spatial feature map by using both average-pooling and max-pooling operations, the average-pooled feature F_{avg}^{c} and the max-pooled feature F_{max}^{c} of spatial context descriptors are generated. Both descriptors are forwarded to a shared network to produce channel attention map $M_c \in R^{C\times1\times1}$. Here C is the number of channels. The shared network is composed of multi-layer perceptron (MLP) with one hidden layer.

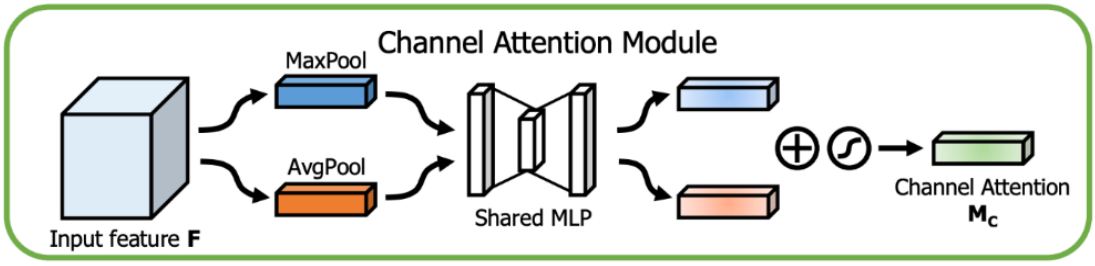


Fig. 2.12. Channel Attention Module (Image source: [30])

To reduce parameter overhead, the hidden activation size is set to $R^{(C/r)\times1\times1}$, where r is the reduction ratio.

Layer attention module.

The feature in high-level layers contains global semantic information, this information can be used as a guide for low-level features to filter the positioning details of different categories. In PAN [39], it's layer attention module design "global attention upsample" has been demonstrated as Figure 2.13.

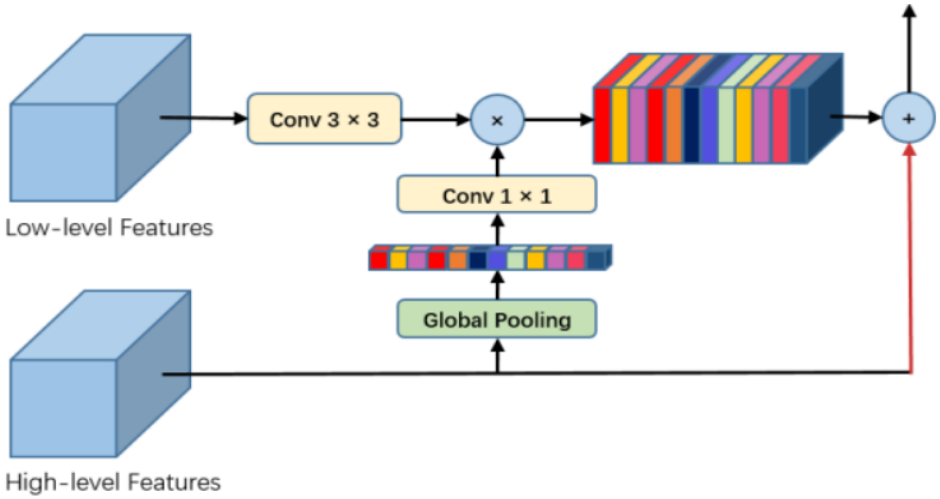


Fig. 2.13. global attention upsample module (Image source: [32])

Mixed attention.

The attention module of different domain often be used together to enhance the deep neural network and improve the performance of downstream tasks. For example, the CBAM [30] combined the channel and spatial domain attention module to enhance the convolution network. And the DANet [40] utilizes both the spatial and channel domain attention to improve the scene segmentation.

Self-attention.

In addition to the various domain related attention module design, the self-attention which improves the attention mechanism by reducing the dependence on external information and captures the internal correlation of data or features has also been widely noticed and studied. In the neural network, the image features are obtained through the linear operation of the convolutional layer. In order to increase the receptive field, convolutional layers stacking is usually adopted, and the effect is limited. At the same time, many tasks of computer vision affect the final performance due to insufficient semantic information. Self-attention obtains larger receptive field and contextual information by capturing global information.

Self-attention [33] has made great progress in sequence models; additionally, context information is critical for many visual tasks, such as semantic segmentation

and target detection. The self-attention mechanism provides an effective modelling method for capturing global context information through triples of (key, query, value). The structure is shown in Figure. 2.14 as below:

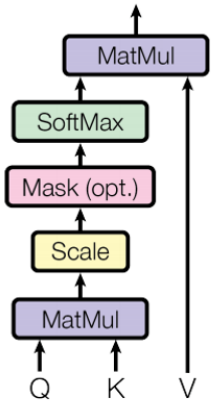


Fig. 2.14. Self-attention (Image source: [33])

For self-attention, the three matrices Q(Query), K(Key), and V(Value) all come from the same input. First, we need to calculate the dot product between Q and K, and then in order to prevent the result from being too large, it will be divided by a scale $\sqrt{d_k}$, where d_k is the dimension of Query and Key. Then use the Softmax operation to normalize the result to a probability distribution, and then multiply it by the matrix V to get the expression of the weight summation. This operation can be expressed as equation 2.8.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2.8)$$

Attention is all you need [34] is the first work that proposed the use of self-attention mechanism to replace cyclic neural networks in sequence models and achieved great success. Non-local neural network [58] is a typical network of utilizing the self-attention in CV. Figure 2.15 shows the basic structure of self-attention of non-local network, where the input $\boldsymbol{X}$ is the feature map obtained by the basic deep convolutional networks, such as ResNet, Xception, etc. The output $\boldsymbol{Z}$ is the saliency map. These basic deep convolutional networks are called backbone. The self-attention structure

contains three branches, the θ, φ and g corresponding to the query, key and value, respectively. The $\otimes$ in the figure denotes matrix multiplication. The calculation is usually divided into three steps.

- The first step is to calculate the similarity between Query and each Key to get the weights. Usually the dot product, splicing and perceptron are used for similarity calculation.

- The second step is to normalize these weights using a softmax function.

- In the third step, the weights and the corresponding key Values are weighted and summed to obtain the final attention.

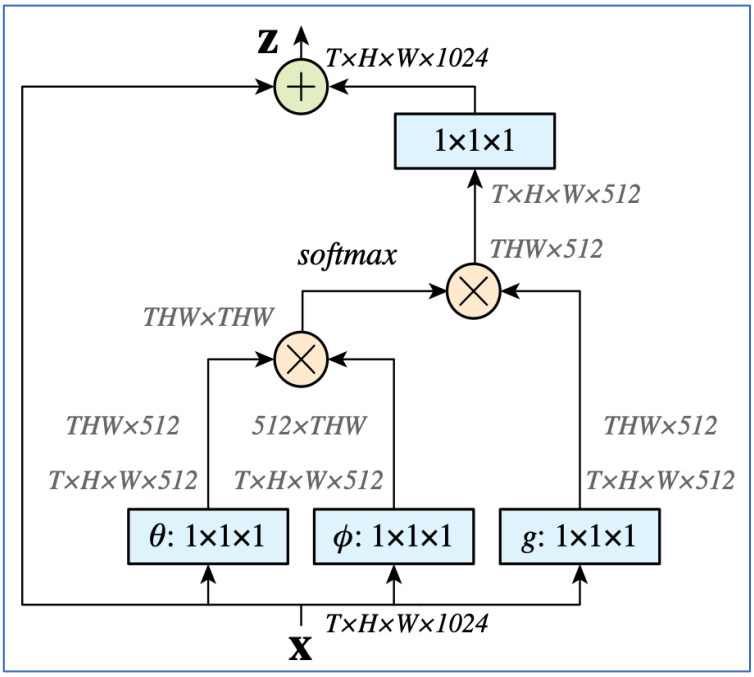


Fig. 2.15 Self-attention structure (Image source: [58])

2.3.3 The proposed pixel level attention for segmentation

Accurate salient object segmentation needs to have a good understanding of both the global semantics of the image and the local contextual information.The network's ability to understand and utilize contextual information is critical to achieving accurate

image segmentation. Increasing the network receptive field by consecutive down sampling is a straightforward way to capture contextual clues for dense prediction. However, the receptive field is restricted by the number of down-sampling operations performed, as performing down-sampling can cause spatial information loss. Inspired by the methods of SPP [36] and ASPP [37] and the approach to locate regional objects by observing neighboring contexts, we propose the PA-ASPP for extracting locally weighted, multiscale contextual information. The PA-ASPP feeds the input into two branches, one is to use the pyramid feature aggregation (top branch) for providing pixel-level attention and the other one is to use the ASPP (bottom branch) for providing rich multi-scale features and large receptive fields. The pyramid feature aggregation encodes the multi-level features from three different scales like U-Net [44]. This feature fusion incorporates neighboring scales of context features more precisely to produce better pixel-level attention.

The obtained attention is up-sampled to the same size of output of ASPP and used to weight the multi-scale features in pixel-level for selecting contextual information. The ASPP enriches contextual information by concatenating features obtained from one 1×1 convolution, three 3×3 convolutions with dilation rates equal to (6, 12, 18), and one global average pooling layer and then passing through another 1×1 convolution to modify channel dimensions of multi-scale features before multiplying the attention maps. Finally, we use a scale-aware selection (SAS) module to locally weight multi-scale contextual features. More specifically, we first use a squeeze-and-excitation module [31] to select the key channels from multi-scale features. Then, the selected features are highlighted pixel- wisely by attention maps from Pyramid Attention Module (PAM). The SAS can capture important contexts of ASPP for the accurate and consistent dense prediction.

2.3.3.1 Pyramid Attention Atrous Spatial Pyramid Pooling (PA-ASPP)

Before introducing PA-ASPP, we first introduce Atrous Spatial Pyramid Pooling (ASPP) as shown in Figure 2.16. The ASPP module is widely used in semantic segmentation tasks due to its effective ability to extract multi-scale context information. ASPP adopts multiple atrous convolutions with different dilation rates in series or

parallel to obtain multi-scale context information to ASPP uses multiple porous convolutions with different expansion rates in series or in parallel to obtain multi-scale context information to solve the problem of the scale diversity of segmented objects.

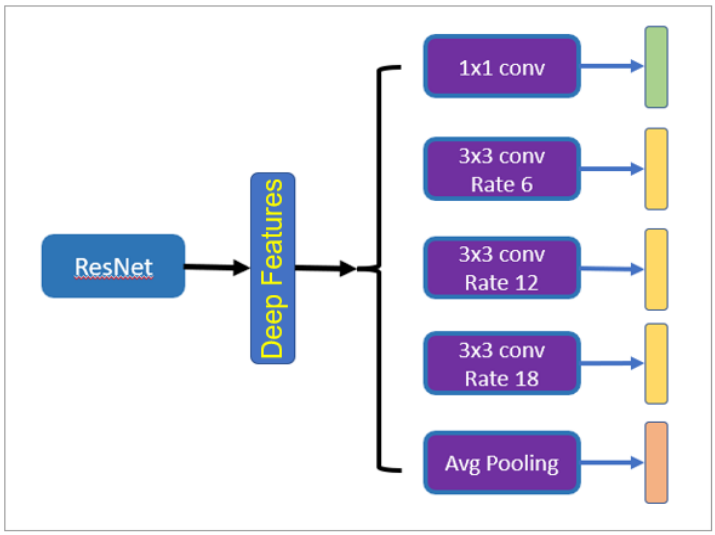


Fig. 2.16 Structure of ASPP

In the ASPP module, dilated convolution is a kind of sparse calculation, which may cause grid artifacts [30], the pyramid pooling module proposed in PSPNet may lose pixel-level localization information. PA-ASPP is introduced to alleviate this problem by providing pixel level attention.

The PA-ASPP is designed to aggregate the global-to-local context information effectively. The PA sub-module (Figure 2.17) generates the pixel-attention map, which is used for locally weighted the multiscale contextual information. PA-ASPP feeds the input into two branches, one is to the pyramid feature aggregation (top branch) for providing pixel-level attention and the other one is to the ASPP (bottom branch) for providing rich multi-scale features and large receptive fields. The pyramid feature aggregation encodes the multi-level features from three different scales like U-Net [44]. This feature fusion incorporates neighbouring scales of context features more precisely to produce better pixel-level attention.

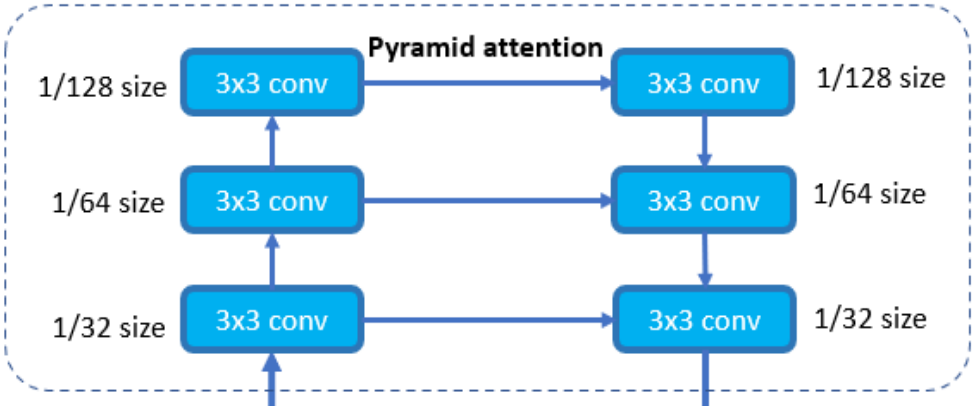


Fig. 2.17 Structure of Pyramid Attention

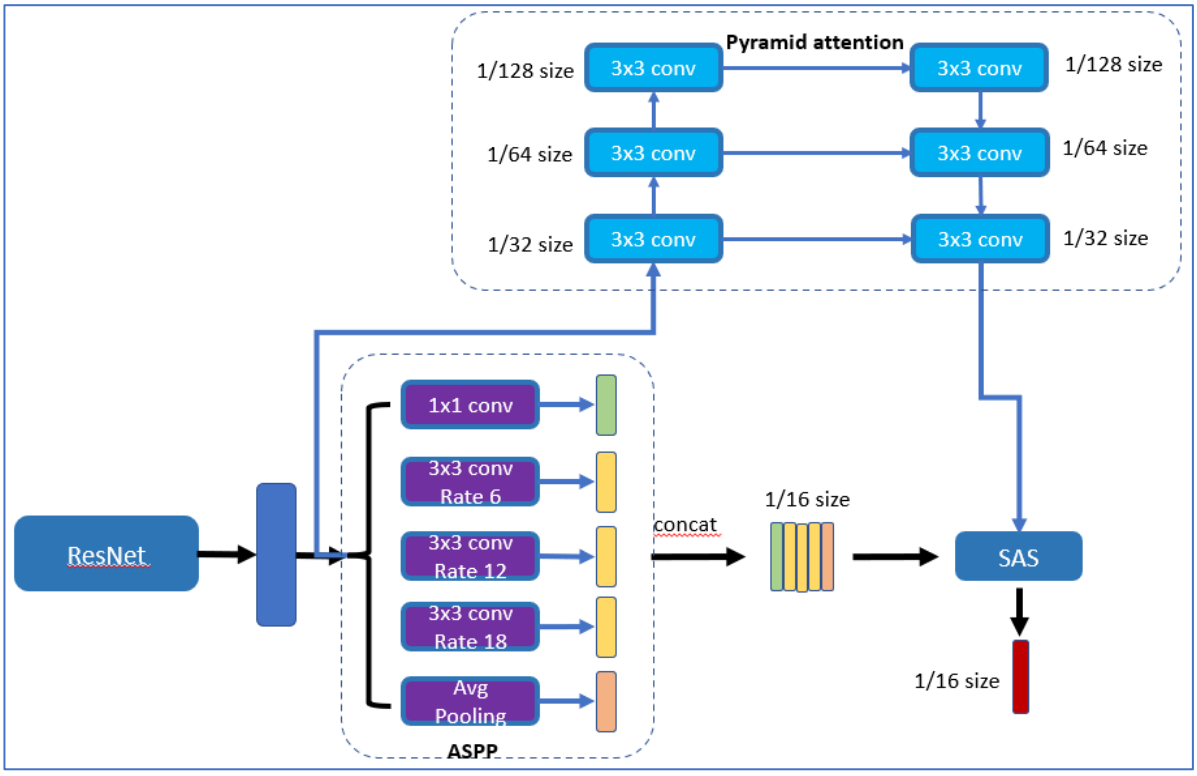


Fig. 2.18 Structure of PA-ASPP

2.3.3.2 Scale-aware selection (SAS)

Scale-aware selection (SAS) module as shown in Figure 2.19 is designed to locally weight multi-scale contextual features. More specifically, we first use a squeeze-and-excitation module [31] to select the key channels from multi-scale features. Then, the selected features are highlighted pixel-wisely by attention maps from Pyramid

Attention Module (PAM). The SAS can capture important contexts of ASPP for the accurate and consistent dense prediction.

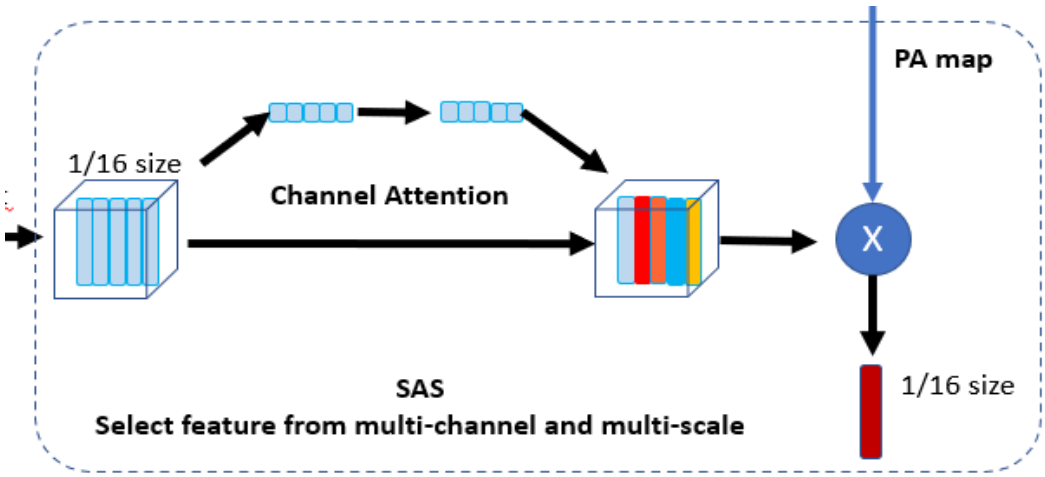


Fig. 2.19 Scale Aware Selection [31]

2.3.4 The Proposed Deep Segment Network (DSNet)

The PA-SAPP and SAS attention module has been integrated with the Resnet and formed a new Deep Segment Network (DSNet). DSNet inspired by [23], as shown in Figure 2.20. The DSNet takes in the patch images, extract multi-level features with dilated ResNet-101 [24]. (Dilation size = 2 at the last residual module), result in the probability map of the metastasis with the 1/16 resolution of input images and extracts locally weighted, multiscale contextual information with the proposed pyramidal attentional ASPP (PA-ASPP) as shown in Figure 2.18.

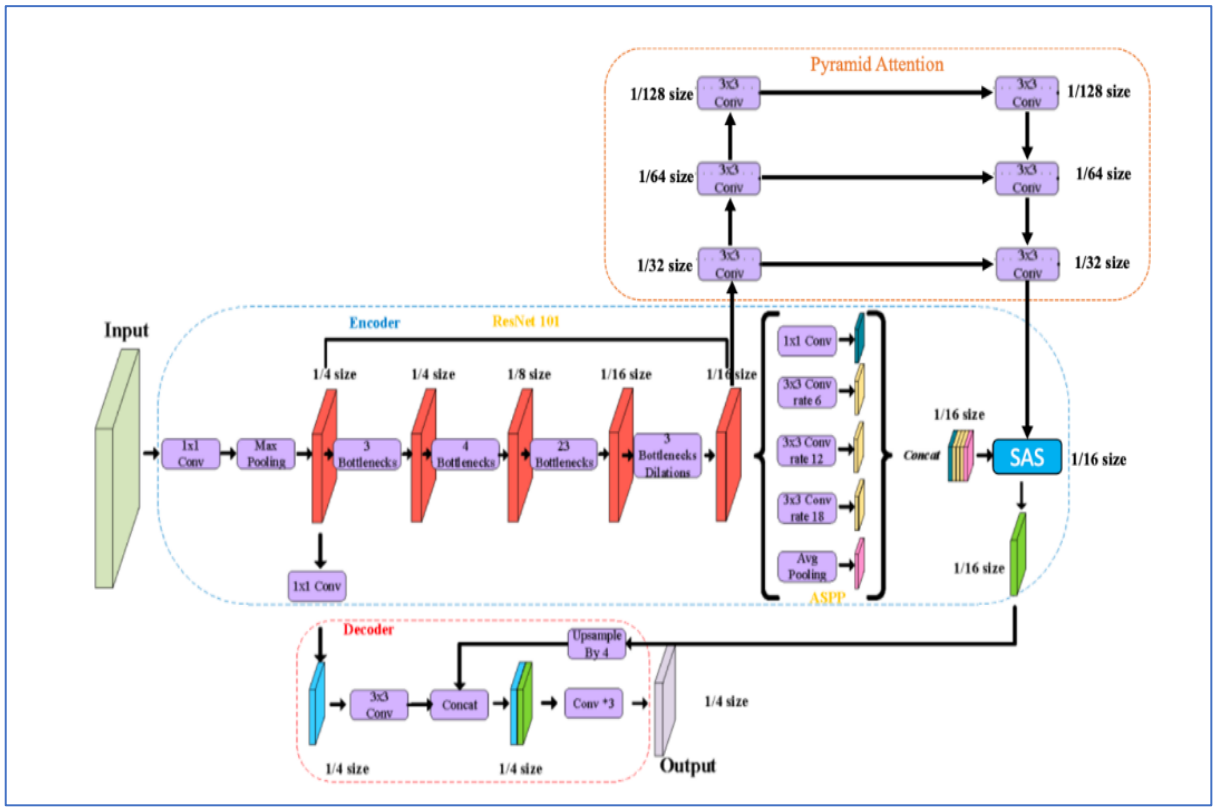


Fig. 2.20. Deep Segmentation Network (DSNet) for saliency object localization.

The encoder part is the dilated ResNet-101 with an ASPP-like contextual structure. The pyramid attention module (PAM) generates attention maps for feature enhancement. the scale-aware selection (SAS) module selects the contextual information. Finally, a skip-connection of low-layers for region localization.

2.3.5 Conclusion

In the chapter, we reviewed the different domain attention mechanism, and then introduced the proposed pixel level attention mechanism of PA-ASPP and Scale-aware selection SAS module. We propose an attention mechanism which can capture important contexts and semantics for the accurate and consistent dense prediction. And based on the proposed attention, designed a deep segmentation model for salient object localization. The effect of the designed attention and the performance of DSNet will be evaluated in practical applications of medical image analysis in Chapter 3.

3 Visual attention in medical imaging applications

In the previous chapter, we have introduced current research status of visual attention for saliency detection and also introduced the proposed semantic attention and attention module in our research. In this chapter, we will introduce a systematic solution which leverage the proposed saliency attention mod in medical image based disease diagnosis and discuss the significance and usefulness of research work on visual attention for practical applications.

3.1 AI-aided diagnosis for medical image analytics

3.1.1 Status of AI-aided diagnosis for medical image analysis

Medical image analysis is an important part of medical diagnosis and treatment. However, there are existed some pain points in the industry, such as the repetitive work, heavy workload and high rate of the misdiagnosis and missed diagnosis. With an overwhelming medical image data volume generated daily for clinical diagnosis, these pain point problems become more serious. The problem is further complicated due to the lack of radiologist / pathologist, rapid increase in workload and rise in imaging complexity. The emergence of artificial intelligence in medical imaging can help to solve these industry pain points, thus attract the attention of the medical industry. The medical image analysis is in great need of automated and computer-aided processing. To address the above issues, a lot of research has been done in both academic and industrial fields, trying to meet the high standard requirements for clinical use.

Nowadays, many scenarios of AI in clinical applications have emerged, which includes:

1) Lesion detection: helping the doctor to early detect the disease, reduce the misdiagnosis. The AI-aided analysis has been widely used, for example, in radiology for the diagnosis of lung cancer and bone fractures, in pathology for breast and cervical cancer screening, and in endoscopy for the lesion detection of digestive tract.

2) Quantitative analysis: providing technical support for precision medicine. Usually, the lesion feature extraction and automatic segmentation related AI assisted technologies are used.
3) Disease diagnosis: Providing accurate judgments of the benign and malignant of tumors, assisting doctors in grading the disease, saving doctors' diagnosis time, and improving the accuracy of diagnosis.
4) Treatment assessment: Providing assistance and advice on the preoperative planning, prognostic treatment, treatment efficiency assessment, etc.
5) Navigation: AI-assisted navigation also plays an important role in the field of surgical robotics.

3.1.2 Major challenges

From a technical point of view, the main challenges of CAD for medical image analysis are reflected in two aspects, the accuracy, and the processing speed.

Accuracy is often affected by data issues. we all know that deep learning-based algorithms require a large amount of precisely annotated data for training to obtain satisfactory model accuracy, especially in the medical field, the requirement is even higher. However, in the field of medical imaging, there are still some data issues existed. First, the amount of annotated data is not large enough, second, there is a lack of precisely annotated data, and third, the data are not evenly distributed. And the annotation itself has the incomplete, inexact, and inaccurate issues. Take the pathology image labelling for abnormal detection as an example. Figure 3.1 shows the issue of incomplete annotation. The pathologist does not label every abnormal cell in the image, which would be very labour intensive and time consuming. Usually, pathologists choose to label a few typical cells. And some time, the pathologist does not label the objects, only give a label to the image if abnormal is found in the image as shown in Figure 3.2. And some time there are some inaccurate labelling existed in the data as shown in Figure 3.3. All these data and annotation issues will affect the accuracy of the model seriously and need to find ways to solve.

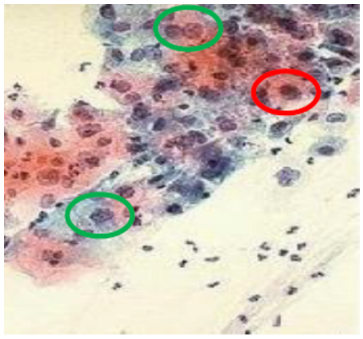


Fig. 3.1 Incomplete annotation (green circle: correct annotation; red circle: missed annotation)

normal ***abnormal***

Fig. 3.2 Inexact annotation (only image level label, no object level label)

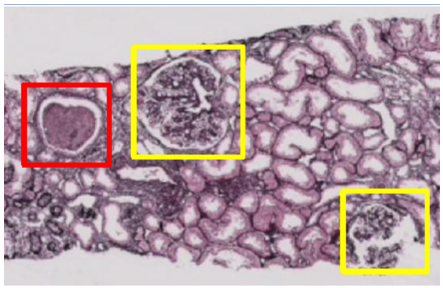


Fig. 3.3 Inaccurate annotation (label in red is incorrect)

In addition to the accuracy, processing speed is very critical in medical image processing. On one hand, it's very important for technologies to ensure the timely and appropriate emergency medical treatment for people who suffer the diseases such as trauma, stroke. On the other hand, medical imaging data occupies around 80% of medical information among the clinical data and is the most important source of information for disease screening and diagnosis, the huge amount data need to be processed as fast as possible to assist the clinical diagnosis. Processing speed is

often affected by the computing resources, the complexity of the problem to be solved and the algorithm model design. In medical image process, the amount of data to be processed are very high and the problem is complex, in the case of determined computing resource, the design of model is critical to the processing speed.

3.1.3 Previous works

Deep learning methods with deep feature extraction and attention mechanism strategies have been shown to be important in different industrial applications in terms of improving accuracy and processing speed, and in medical image processing field, the attention for saliency detection technology also play some effective role. Such as the synthetic medical image dataset generation [56, 58, 59], explainable AI [60, 61] in medical image analysis and lesion detection [62] and segmentation [63] for diagnosis.

3.2 Visual attention system for breast cancer detection

3.2.1 Background introduction of breast cancer detection

Medical image segmentation is a fundamental and challenge task in many computer-aided diagnosis (CAD) tasks. These CADs conduct rapid analysis and understanding of large amount of medical data to reduce the doctor's interpretation time, the segmentation algorithm is the fundamental function for the quantitative analysis and risk evaluation. There are multiple modality medical images, such as the X-ray, MR and CT. In our work, we focus on breast cancer detection based on pathological image analysis.

Digital pathology image.

The pathology glass slices are digitized by digital scanning. After digitization, a huge digital image (Gigapixel Pathology Images) will be generated usually the size is about 100,000 * 1000,000 pixels, which is called whole slide image (WSI). WSI has multi-resolution pyramid structure, it consists of multiple slide images and these slide images have different scanning magnifications. Mimicking the observation of the human eye under different magnifications of the microscope, digital slides are usually

digitized and stored at 1x, 5x, 10x, .20x and 40x magnifications as shown in Figure 3.4. The average size of each WSI is over 1G or even 3-5G.

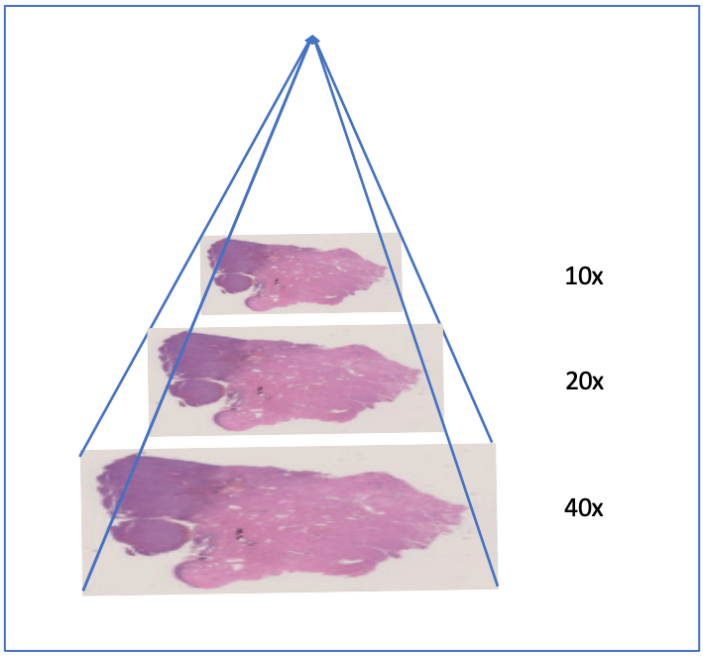


Fig. 3.4. The multi-resolution pyramid structure of whole-slide image

Breast cancer.

Breast cancer is the most common cancer among women. According to the statistics from International Cancer Research Center (ICRC) of the World Health Organization (WHO), breast cancer impacts 2.1 million each year, and causes the greatest number of cancer-related deaths among women. In 2018, it is estimated that 627,000 women died from breast cancer - that is approximately 15% of all cancer deaths among women. Globally, the incidence rates vary widely, and the age-standardized incidence rate is as high as 99.4 per 100,000 in North America. The incident rates in Eastern Europe, Southern America, Southern Africa, and Western Asia are slightly lower, however they are increasing at the same time. The improvement of survival rates can be mainly achieved through early detection and diagnosis, which require a large proportion of women to seek medical treatments, appropriate diagnosis, and treatment facilities. Early detection and classification of breast cancer metastases in lymph nodes play an important role in appropriate

diagnosis and treatment facilities. Lymph nodes are small glands that filter lymph, the fluid that circulates through the lymphatic system, as shown in Figure 3.5.

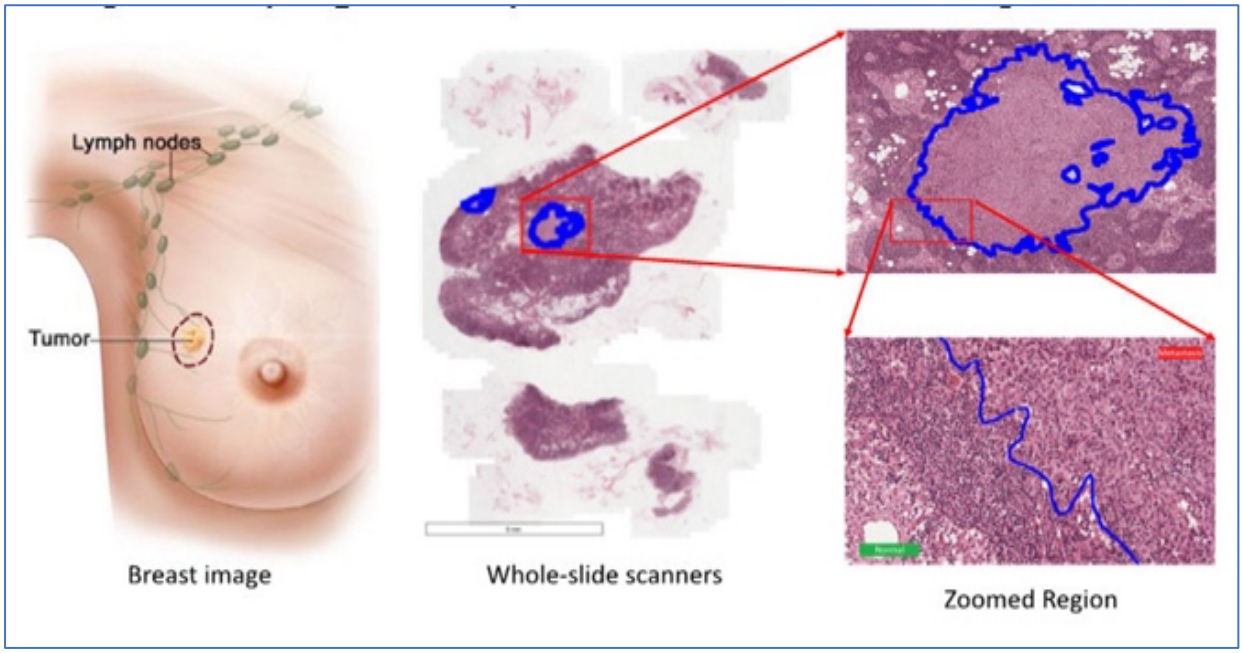


Fig. 3.5. Example of a typical metastatic region. Whole-slide images (WSI) are generally stored in a multi-resolution pyramid structure. Each image in the pyramid is stored as a series of tiles to facilitate rapid retrieval of subregions of the image (https://camelyon17.grand-challenge. org/Background/). The regions of metastases have very similar appearances to the normal regions of scanned sections.

Generally, automatic diagnosis of the presence of metastases in lymph nodes has therapeutic implications for breast cancer patients. Detection and classification of breast cancer metastases have high clinical relevance, especially in whole-slide images of histological lymph node sections. Fast early detection leads to huge improvement of patients' survival rate. However, currently pathologists mainly detect the metastases with microscopic assessments. This diagnosis procedure is extremely laborious and prone to inevitable missed diagnoses. Therefore, automated, accurate patient-level classification would hold great promise to reduce the pathologist's workload while also reduce the subjectivity of diagnosis. In the following chapters, the Deep Regional Metastases Segmentation (DRMS) framework for the patient-level lymph node status classification will be introduced.

Currently, the UICC TNM classification system is an internationally recognized means of classifying the extent of cancer spread in patients with solid tumors. It is one

of the most important tools to help clinicians select suitable treatment options and obtain prognostic indications. In breast cancer, UICC TNM staging takes into account the size of the tumor (T-stage), whether the cancer has spread to the regional lymph nodes (N-stage), and whether the tumor has metastasised to other parts of the body (M-stage). Since the histological assessment of lymph node metastases is an essential part of UICC TNM classification, in this paper we will focus on the pathologic N-stage (pN-stage).

Automatically detecting the regional metastases, especially from the whole-slide images, is a very challenging task due to the low contrast and similar structures between metastasised regions and various adjacent tissues or organs, as shown in Figure 3.5. Moreover, the interference from other tumors makes the metastasised regions even more difficult to be separated from the background. In some cases, the metastasised regions are extremely small, which further complicates the task.

3.2.2 Deep Regional Metastases Segmentation (DRMS) framework

We propose an visual attention based deep regional metastases segmentation (DRMS) framework to achieve prediction of pN stage scores from whole slide histopathology images of patients. The proposed framework consists of four key modules: a region of interest (ROI) extraction module, a patch-level regional metastases localization module, a slide-level metastasis detection module and a patient-level lymph node classification module.

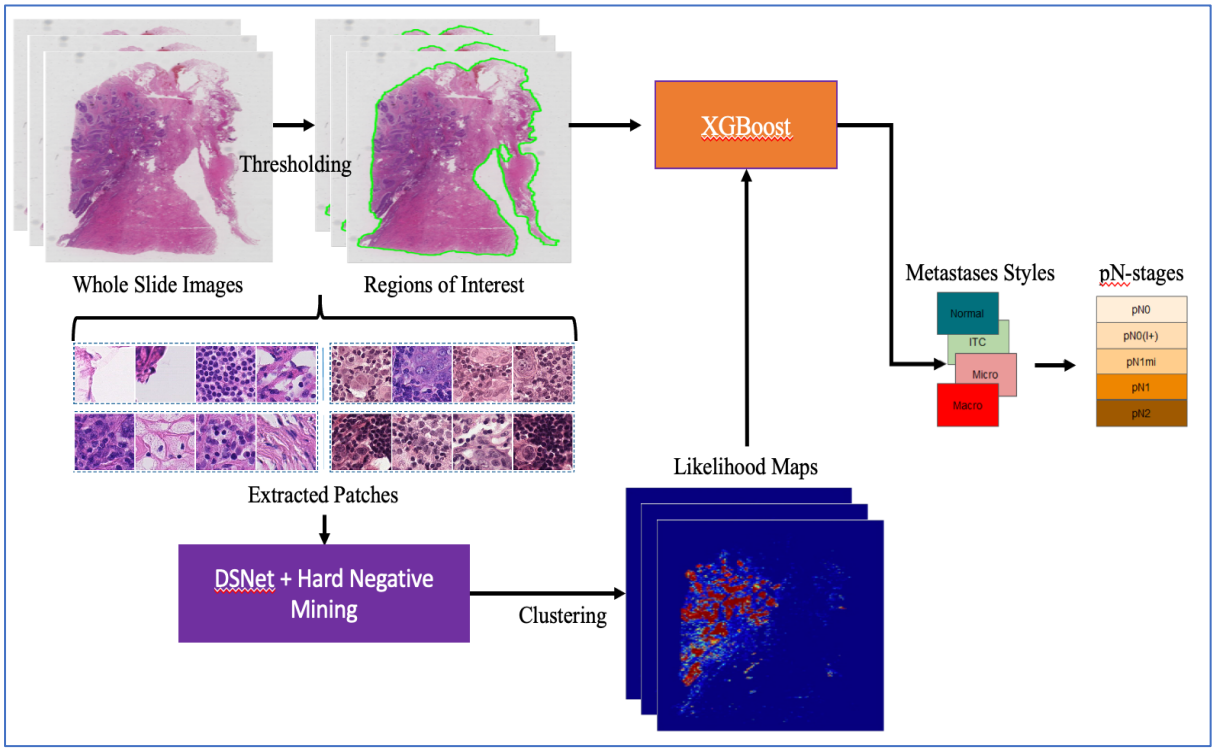


Fig. 3.6. The overall procedure of our proposed framework.

Multi-grained procedure is designed for detecting cancer metastases on gigapixel pathology images. With multi-grained analysis, the proposed method can capture rich information from pathology images.

Figure 3.6 shows the proposed Deep Regional Metastases Segmentation framework. First, the ROI extraction module generates candidate tissue regions by splitting the whole slide images (WSIs) into smaller patches; then, a Deep Segmentation Network (DSNet) is proposed to detect the regional metastases in patch-level; After that, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [46] is adopted to cluster the slide level feature, generate the likelihood maps, and predict the cancer metastases within extracted ROIs. Then, the slide-level metastasis detection module adopts the DBSCAN to fuse the predicted scores extracted from ROIs to build a slide-level lymph node prediction [3]. Finally, the patient-level pN-stages are determined by aggregating each individual slide-level prediction. In combination with the above techniques, the framework can make better use of the

multi-grained information in histological lymph node section of whole-slice images. Experiments on large-scale clinical datasets (e.g., CAMELYON17) demonstrate that our method delivers advanced performance and provides consistent and accurate metastasis detection in clinical trials.

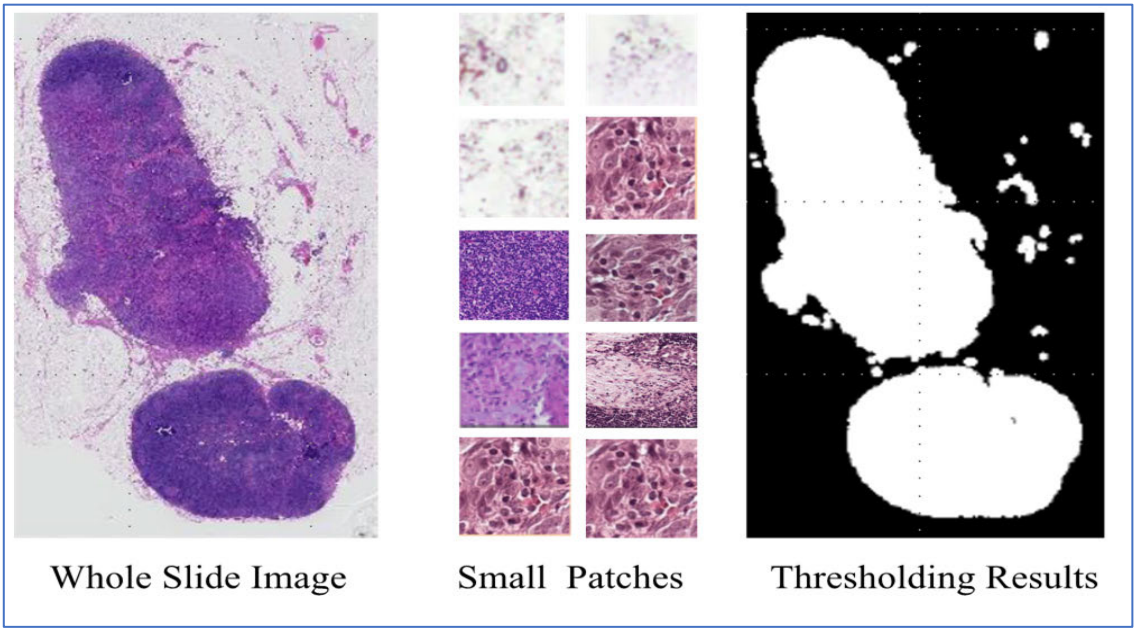


Fig. 3.7 From whole slide images to patches.

ROI Extraction.

For histopathology images, a typical whole slide image is approximately 200,000 × 100,000 pixels on the highest resolution level. If we deal with such large images directly, enormous computation is required because of the huge size of the slide. Besides, most of the regions have no useful information for metastases detection. Thus, directly dividing WSIs into patches is not appropriate for model training, as shown in Figure 3.7. The Otsu threshold [15] method is widely used in recent studies for ROI extractions from the WSI because of its simple calculation and the segmentation is not affected by image brightness and contrast. In this work, to speed up the processing, we also adopt Otsu's adaptive thresholding technique on the low-resolution version of the WSIs to remove most of the non-tissue background quickly. Following [22], we convert RGB to gray from 32-times down-sampled WSI and then extract tissue regions by Otsu threshold method. We observed that metastasis regions

are usually located at the edge of the tissue regions. Therefore, careful tissue region extraction method is needed. After the operation is done, holes and small points, which will influence feature extraction operation, still exist. Following previous works, morphological algorithms are operated to fill holes and clear isolated points. Then, we generate the patches through random grid sampling. Examples of the resulting patches are shown in Fig. 3.7. We can see that although most examples have texture patterns, it is difficult to distinguish between normal or metastatic patches.

3.2.3 Patch-level metastasis localization

The patch-level localization aims to predict the metastases regions at coarse granularities, i.e., to find the saliency regions, which represent the metastasis locations. To achieve this goal, we opt to a deep segmentation network (DSNet) shown in Figure 2. 20.

To train the proposed network, we adopt the input resolution of 512 × 512 to capture more detailed information. To increase the variation in the training set, we perform extensive data augmentation including random flipping, scaling, rotation over angles between 0 and 360, and cropping. Besides, hard negative mining is conducted during the training procedure to improve the training dataset. It means that for each training epoch, the model inferences the whole patches of the slide and the patches whose intersection over union (IoU) with the annotated mask are less than 0.95 are selected as the negative training data. Hard negative mining significantly improves the models' discrimination ability.

Tumor regions cover a very small proportion of pixels in WSIs, thereby leading to class imbalance. This issue was circumvented by training the network to minimize a hybrid loss function. In this work, the hybrid cost function comprised of a class-weighted cross-entropy loss and a loss function based on the Dice overlap coefficient. The Dice coefficient is an overlap metric used for assessing the quality of segmentation maps. The Dice loss is a differentiable function that approximates to Dice coefficient, it uses the predicted posterior probability map and ground truth binary image, as defined in equation 3.1. The weighted cross-entropy loss is defined in

equation 3.2. In the equations, pi is the predicted posterior probability map, and gi is the ground truth image.

$$L_{dice} = 1 - 2\frac{\sum_l^L w_l \sum_n g_{ln} p_{ln}}{\sum_l^L w_l \sum_n g_{ln} + p_{ln}} \qquad (3.1)$$

$$L_{wce} = -\frac{1}{N}\sum_1^N (wr_n \log(p_n) + (1 - r_n)\log(1 - p_n)) \qquad (3.2)$$

The total loss is defined as a linear combination of the two loss components as defined in equation 3.3.

$$L = \alpha L_{dice} + \beta L_{wce} \qquad (3.3)$$

The proposed networks are trained by minimizing the total loss. α and β are empirically assigned to the individual loss components. In this work we set α=β=0.5.

3.2.4 Slide-Level Metastases Detection

To capture the slide-level information, we integrate the patch information to improve the spatial consistency. Following the work in [47], we introduce the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to group together small metastases areas which were in close proximity.

DBSCAN is a density-based spatial clustering that solves applications with noise. It is able to find arbitrary shaped clusters and clusters with noise (i.e., outliers). Due to the pros of the DBSCAN showed below, it is suitable for grouping the small metastases areas. Figure 3.9 shows an example of metastases probability map grouping in slice level by DBSCAN.

- DBSCAN does not require to specify number of clusters beforehand.
- DBSCAN performs well with arbitrary shapes clusters.
- DBSCAN is robust to outliers and able to detect the outliers.

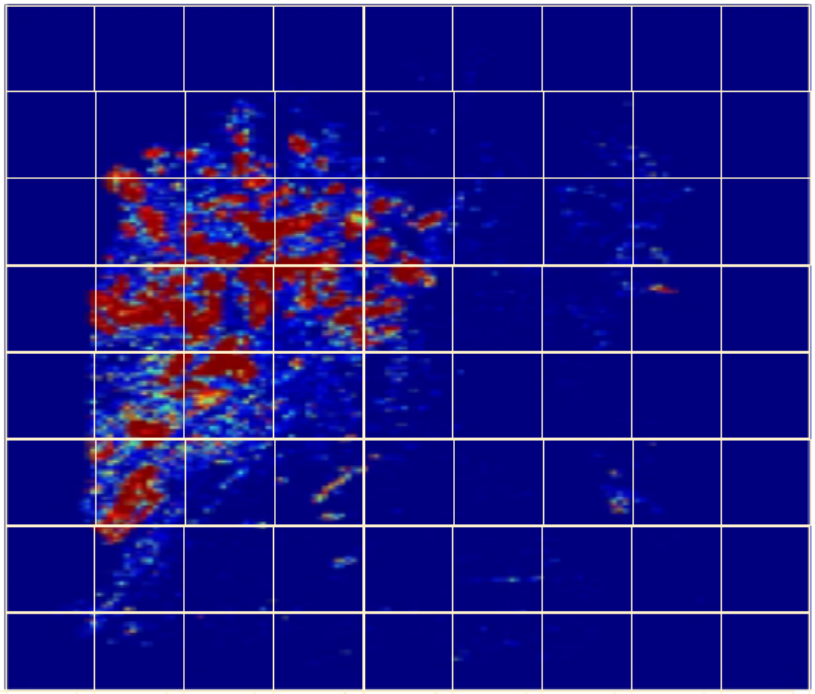


Fig. 3.9. Probability map of metastases grouping in slice level by DBSCAN

3.2.5 Patient-level Classification

After slide-level metastases detection, we can realize the patient-level classification based on the corresponding metastases regions.

The task is to determine a pN-stage for each patient in the test dataset. To compose a pN-stage, the number of positive lymph nodes (i.e., nodes with a metastasis) need to be counted. According to the size of the tumor, there are three types of positive lymph nodes in clinic:

- Macro-metastases: Metastases greater than 2.0 mm.
- Micro-metastases: Metastases greater than 0.2mmor more than 200 cells, but smaller than 2.0 mm.
- Isolated Tumor Cell (ITC): Single tumor cells or a cluster of tumor cells smaller than 0.2 mm or less than 200 cells.

Although lymph nodes containing only ITCs are not counted as positive lymph nodes, if no macro-metastases or micro-metastases are found in the patient's lymph nodes, the pathologist must report ITC.

Following the previous work, we implement a simplified version of the pN-staging system for breast cancer, that can automatically determine the following applied pN staging for each patient.

- pN0: No micro-metastases or macro-metastases or ITCs found.
- pN0(i+): Only ITCs found.
- pN1mi: Micro-metastases found, but no macro-metastases found.
- pN1: Metastases found in 1–3 lymph nodes, of which at least one is a macro-metastasis.
- pN2: Metastases found in 4–9 lymph nodes, of which at least one is a macro-metastasis.

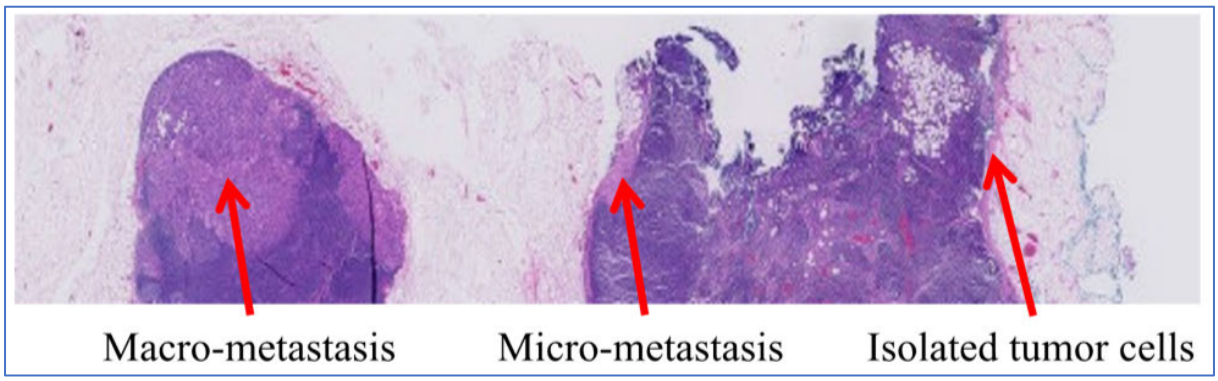


Fig. 3.10. Representative samples of the different sizes of breast cancer metastases in sentinel lymph nodes.

In order to perform the patient-level classification, we design and extract the meaningful features from slides' probability maps. More specifically, to train a pN-stage classifier, we structure the features by converting each whole slide probability map into a feature vector by regulating it with 5 threshold values, probability maps are generated by these threshold values and formed into a feature vector. Then according to the morphological and geometrical information, we extract 13 types of features from each probability map, 65 features in total.

After feature vectors are extracted, we use XGBoost to classify lymph nodes into four classes (Normal, ITC, Micro, Macro). Since the total number of WSIs is rather small, it is important to prevent overfitting. Thus, K-fold cross validation and other hyper-parameters, such as subsample or max depth, are adopted for preventing overfitting. Firstly, grid-search algorithm is used to search for the best hyper-parameters. Then, the XGBoost model is trained with ten-fold cross validation. Finally, the best model is selected and the 100 patients in the CAMELYON17 test set are predicted. The patient's pN-stage is graded according to the given rules.

Below gives a simple introduction on the XGBoost algorithm and grid-search algorithm.

- XGBoost [55] is the best algorithm among gradient boosting decision tree (GBDT). It is an open-source software library which provides a regularizing gradient boosting framework for different languages, such as C++, Java, Python.

  XGBoost has been widely used because of its advantages. Firstly, XGBoost has a regular term in the cost function, which helps control the complexity of the model. The regular term reduces the variance of the model, which makes the learned model simpler and prevents overfitting. Second, XGBoost supports parallel processing in terms of feature granularity. Again, XGBoost is very flexible, it supports user-defined objective function and evaluation function. In addition, for samples with missing features values, XGBoost can automatically learn its splitting direction. Finally, XGBoost subtree building, and pruning do not easily fall into local optimal solutions.

  XGBoost certainly has its own drawbacks and shortcomings. First, XGBoost has too many parameters of the algorithm, complicated to adjust the parameters. Second XGBoost algorithm is only suitable for processing structured feature data, but not good for unstructured data with heavy tasks such as target detection in images. Finally, it is not suitable for processing ultra-high-dimensional feature data.
- The grid search algorithm in short is a parameter tuning method. By exhaustive enumeration, a model is trained for each possible combination of parameters, and

then model evaluation is performed to find the set of parameter values when the model is optimal based on the evaluation results.

High precision classification and fast processing speed are crucial in digital pathology image analysis. In patient-level classification, the data has been structured. It is reasonable to use XGBoost.

3.2.6 Performance analysis

Datasets.

For training patch-level metastasis localization model DSNet, the WSIs from CAMELYON16 are also used as the training dataset. The CAMELYON16 dataset contains 400 WSIs with region annotations for all its metastasis slides. These WSIs are collected from two different medical centers. The CAMELYON17 dataset contains 1000 WSIs with 5 slides per patient: 500 slides are for the train set, the other 500 slides for the test set. The WSIs are collected from five different medical centers.

Typically, different sizes of breast cancer metastases appear in sentinel lymph nodes, training datasets with regional annotations are critical. Since the CAMELYON17 dataset provides only 50 slides with region annotations, these slides belong to different patients, we split 100 patients (500 slides) into 43 patients for the train set, 57 patients for the validation set for hyper-parameter tuning. In detail, if a patient's any slide includes region annotation, we allocate that patient into the train set. For training of the patch-level localization, 400 WSIs from CAMELYON16 dataset and 160 WSIs, where 50 WSIs with region annotation and 110 negative WSIs from CAMELYON17 train set are used.

For training of the slide-level lymph node classifier, we use 285 WSIs from 57 patients from the CAMELYON17 validation set.

The DSNet is evaluated on CAMELYON17 [42] dataset.

Evaluation Metrics.

To evaluate the performance of different methods, we adopt the quadratic weighted Cohen's kappa [41] as the evaluation metric. Given n test patients and m categories (pN-stages), let n_{ij} denote the number of patients with the pNi-stage that were categorized to pNj-stage. Let r_i denote the total number of patients with the pNi-stage and s_j the total number of patients categorized to the pNj-stage. Finally, let w_{ij} denote the disagreement weight associated with the pNi and disagreement weight associated with the pNi and pNj.

The weight matrix is

$$w_{ij} = (i-j)^2, i, j \in [1, \dots, m] \qquad (4)$$

The mean observed degree of disagreement is

$$D_o = \frac{1}{n}\sum_{i}^{m}\sum_{j}^{n} n_{ij} w_{ij} \qquad (5)$$

The mean degree of disagreement expected by chance is

$$D_e = \frac{1}{n^2}\sum_{i}^{m}\sum_{j}^{n} r_i s_j w_{ij} \qquad (6)$$

The weighted kappa is then defined by

$$k_w = \frac{D_e - D_0}{D_e} \qquad (7)$$

The k_w metric ranges from -1 to +1: a negative value indicates lower than chance agreement, zero indicates exact chance agreement, and a positive value indicates better than chance agreement. As pN-stages are ordinal, choosing a quadratic weighted kappa could penalize misclassification more severely when errors are more than one stage apart.

Training Details.

During training and inference, we first perform the thresholding for localizing the ROIs, then extracted 512×512 patches from WSIs at the highest magnification level resolution. We adopt the ImageNet pre-trained ResNet101 [43] (He et al. 2016) as the backbone with dilation convolutions. New layers are initialized with the MSRA method. The deep learning toolbox TensorFlow [29] was used to train the model with a NVIDIA GTX TITAN X GPU. We used the Adam optimization method [30] with a learning rate 1e-4. The network was trained for 4 epochs with a batch size 32. For the post-processing, we thresholded the tumor probability heat map with 5 threshold values, i.e., 0.5, 0.6, 0.7, 0.8, 0.9. Given a heat map, we extracted 13 features including the major axis length of the tumor region (Feature1), maximum probability score (Feature2), average probability score (Feature3), a total area of the tumor region (Feature4), max, mean, variance of eccentricity (Feature5-7), max, mean, variance of orientation (Feature8-10) and max, mean, variance of solidity (Feature11- 13) from the tumor regions. We built a XGBoost classifier to discriminate lymph node classes using extracted features. Finally, each patient's pN-stage was determined by the given rule together with the lymph node slide prediction result.

Comparisons with Other State-of-the-Art Methods.

Using a single segmentation model, our slide-level lymph node classification model and patient-level pN-stage prediction achieved 0.9351 accuracy and 0.9017 quadratic weighted kappa score, respectively. We trained additional models with different data augmentation. Finally, four models were ensembled by averaging probability heat map and reached 0.9410 slide-level accuracy and 0.9632 patient-level quadratic weighted kappa score. Table 3.1 shows the published results on CAMELYON17 testing dataset. Our best performing model gave a Cohen's kappa score of 0.9473, which is better than the second in Open Leaderboard (out of 120 valid submission entries).

Rank	Team	Cohen Kapa	Affiliation	Date
1	shlee	0.9570	Deep Bio Inc.	12 Feb. 2019
2	nicolas.pinchaud1	0.9386	Context Vison	6 Dec. 2019
3	jarveelee3	0.9243	SenseTime Research	28 Feb. 2020
4	mahendrakhened	0.9090	Indian Institute of Technology Madras	6 April 2020
5	ozymandias.watchman	0.9085	42.0	25 March 2019
6	GigapixelAI	0.9058	GigapixelAI	12 Sept. 2019
7	kvougas	0.9052	DeepMed IO	27 Nov. 2019
8	erizz	0.9046	Imsight Medical Tech Inc. and CUHK	10 March 2019
9	jaylee	0.8993	Lunit Inc.	31 July 2017
	ours	0.9473		25 Nov. 2020

Tab 3. 1. Performance comparison with other approaches for automated pN-Staging in CAMELYON17 challenge. The score of Cohen Kapa is from the open public leader board.

According to the brief description of the top-ranked model, it used DeepLabv3+ as the basic framework for patch level segmentation. However, the top-ranked model used the DenseNet with high generalization ability as the backbone of the Encoder. Our proposed DSNet+ adopted Resnet as the backbone and improve the DeepLabv3+ framework by introducing the PA and SAS attentions to enhance the model generalization ability. In the future, we will explore more advanced segmentation frameworks and backbone networks for the dedicated task.

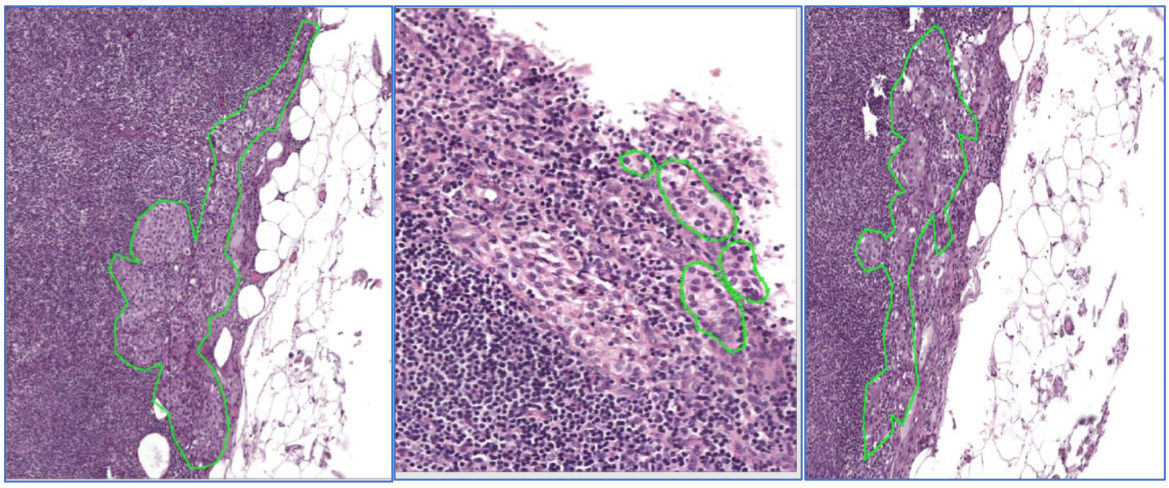


Fig. 3.11 Slide level prediction of our proposed framework in different cases, our method can show good results in continuous regions (a), isolated region (b) and low contrast region (c)

Figure 3.11 shows an example for slide-level predictions. From the results, we can see that our method can achieve outstanding performance in different cases, including continuous regions (a), isolated regions (b) and low-contrast regions (c).

Ablations studies.

Comprehensive experiments are conducted to demonstrate the effectiveness of the proposed method. The 5-fold cross-validation is applied to validate the stability and generalization of networks. It is worth mentioning that all models are trained following the same experimental settings.

- **Effects of the PA-ASPP and SAS**.

  The proposed PA-ASPP is used to aggregate the global-to-local context information. The PPM [38] and ASPP [36] also aggregate multi-scale contextual features for semantic labelling. To verify the effects of our PA-ASPP, we adopt the PSPNet [38] with the ResNet-101 backbone for a fair comparison. We replace the PPM of PSPNet [38] with our PA-ASPP and ASPP [36] modules. The 2-4 rows of Table 3.2 show the quantitative results. With our PA-ASSP module, the Cohen Kappa score improves about 3.3% and 1.6% over the PPM and ASPP, respectively. Visual comparisons of PPM, ASPP and our PA-ASPP are shown in Figure 3.12. It can be observed that our segmentation results constantly contain more accurate and coherent structures. These results demonstrate that our proposed PA-ASPP is more effective than previous methods and can generate more coherent results. Besides, with the SAS, the model can adaptively capture the regions with different scales, which shows the effectiveness.

Methods	KappaScore
PSPNet (PPM) (Zhao et al. 2017)	0.9012
PSPNet (PPM→ ASPP)	0.9188
PSPNet (PPM→PA-ASPP)	0.9341
DeepLabv2 (L.-C. Chen, Papandreou, et al. 2018)	0.9002
DeepLabv3+ (L.-C. Chen, Zhu, et al. 2018)	0.9213
Ours w/o PA	0.8422
Ours w/o SAS	0.9327
Ours	0.9473

Table 3. 2. Ablation results on the camelyon17 validation dataset.

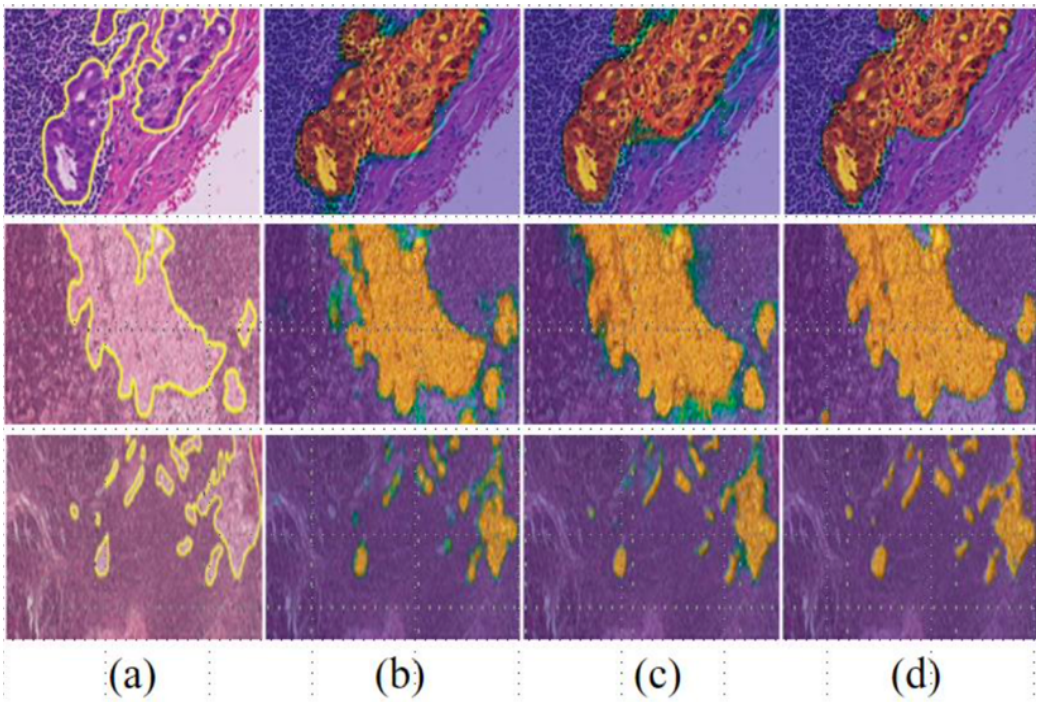


Fig. 3.12. Visual comparisons with different contextual modules. From left to right: Ground truth, PPM, ASPP and our PA-ASPP.

- **Effects of Hard example mining.**

Previous patch extraction scheme from the whole-slide images is based on random uniform sampling. However, this leads to some hard examples being excluded from the training set which resulted in the model's poor performance on such regions of WSI. Therefore, we attempt to solve this issue by hard example mining the poorly

performing regions in WSI and fine-tuning the trained model with this hard-mined set. To cancel the effects of the PA-ASPP module, we apply them to the DeepLabv3+ for direct performance comparison. Tab. 3 shows the quantitative performances. Compared with the original DeepLabv3+, the HEM can consistently improve the classification results. Note that the HEM is a general proposal. It can be used in varied settings for training.

Methods	KappaScore
DeepLabv3+ without HEM	0.9365
DeepLabv3+ with HEM	0.9473

Table 3. 3 Effects of hard example mining (HEM).

- **Effects of the thresholding.**

For predicting patient-level results, we convert each whole-slide probability map into a feature vector with 5 threshold values, i.e., 0.5, 0.6, 0.7, 0.8, 0.9. Based on the regions, we train a lymph node classifier. To verify the effect, we show the visual results in Fig 3.13. As we can see, at a low threshold, the final prediction can include noise for the cancer localization. While at a high threshold, it will miss some key regions. Our multiple thresholds can effectively highlight the most of regions. Figure 3.13 shows the effects with different thresholding values.

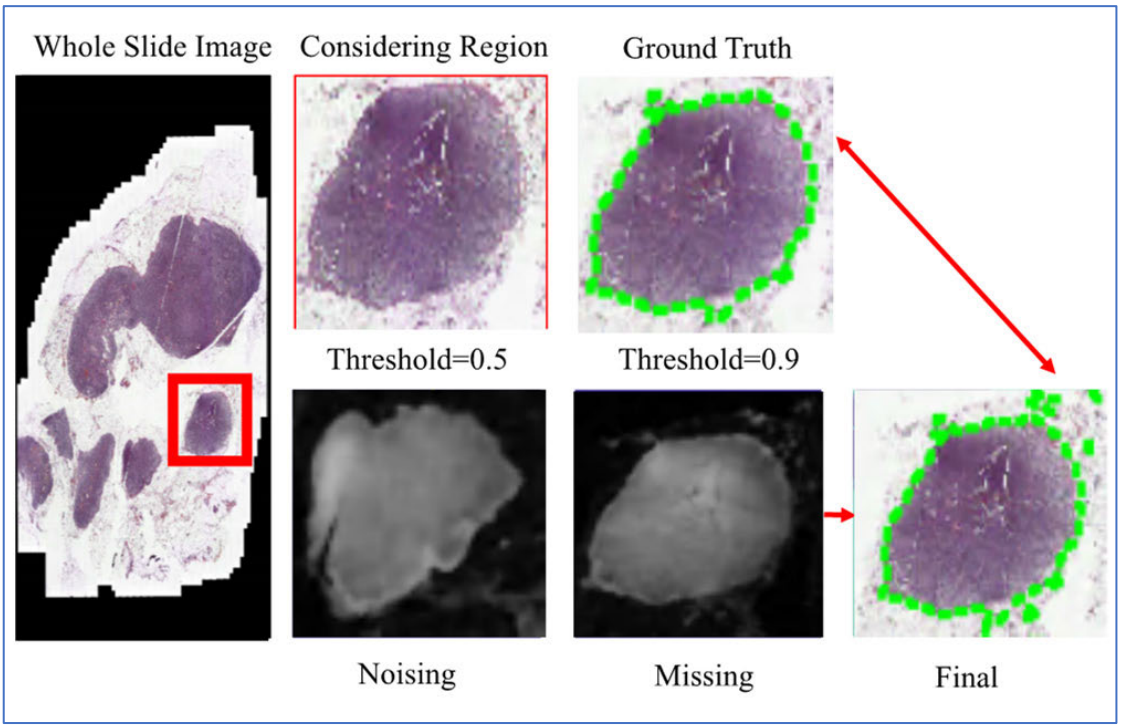


Fig. 3.13. Slide-level predictions of our proposed framework with different thresholding values.

- **Limitations**.

  We address the patient-level lymph node status classification problem by regional metastases segmentation of gigapixel whole-slide images using the divide and conquer strategy. The proposed approach shows superior performance when compared to its competitors. However, our method still shows low accuracy for ITC, and the model size is still very large for real applications. Besides, the patch extraction is based on the Otsu thresholding. In some cases of CAMELYON2017, the complex background is still be included.

3.2.7 Conclusion

In this work, a fully automatic metastases detection framework to predict pN-stage of patient-level lymph node statuses from the whole-slide histopathology images.

Several useful strategies are proposed to improve the regional segmentation accuracy and integrate multi-granularity representations for the final classification. Our method adopts a divide and conquer strategy and allows the model to process the

whole-slide images. More specifically, our method first divides the whole-slide images into tumor-related patches, then performs tumor segmentation on regional patches, and integrates segmented patches to generate the segmentation of the entire WSI. More importantly, we propose a new deep segmentation network (DSNet) with pyramid attentions and scale-aware selection so that the expressive multi-level features can be emphasized to facilitate the feature refinement. Extensive comparative experiments with 5-fold cross-validation on CAMELYON17 demonstrate that our proposed method can significantly outperform several state-of-the-art methods in typical metrics. Due to the characteristics of digital pathological images, it's impossible to manually annotated large quantity qualified datasets for model training, [34] has proposed a weakly supervised deep learning framework to solve the issue. To enable clinical application, we will improve the patch sampling method to generate more valuable training examples; and we will improve the system framework to enable weakly supervised learning and semi-supervised learning. On the other side, according to the review article [35], the applications for DL based pathology analytics for Breast Cancer are still focused on tumor detection, the basic image analysis and at internal validation stage. To enable the advanced clinical applications, we will explore more advanced segmentation frameworks and back- bone networks for the dedicated task, and we will build an end-to-end learning framework to predict the pN phase of the WSI and enable the accurate grading and subtyping clinical applications.

4 Overall conclusion and future work

4.1 Conclusion

The visual attention based saliency detection and visual attention module design in DL are hot research topics in the field of computer vision, and its application is reflected in many CV fields. The latest research results have significantly improved the accuracy and speed of deep learning algorithms. Nevertheless, there are still many challenging issues that need to be solved urgently. Such as how to localize salient objects accurately in complex scenarios; How to make the attention module focus on really useful high-level semantic in different applications. In response to the above problems, in the research work, we focus on the saliency detection, explore the method of the attention based on semantic segmentation and the attention module design for deep learning model, and applied it to medical image analysis field. The main work is summarized as follows:

1. We proposed a novel attention mechanism, semantic attention (SeA) to separate the proposals for saliency object detection. The attention is established based on the observation that regions with high attention should have similarly semantic concepts with salient objects. The SeA takes the high-level semantic features from Faster Region-based Convolutional Neural Network (Faster R-CNN) to assist the proposal selection in images with complex background. The salient proposals are selected according to their semantic attention probabilities.
2. We propose a deep segmentation network (DSNet) for salient object localization, which has been successfully used in patch-level metastases localization for breast cancer detection. The proposed network can aggregate the global-to-local context information of each patch with the PA-ASPP and SAS attention module. Experiments on public benchmarks demonstrate that our approach achieves superior performance and obtains accurate results on metastases detection.
3. We propose a DRMS framework which leveraging the visual attention mechanism to predict pathologic N-stage (pN-stage) from patient's whole slide histopathology images. We propose a multi-grained procedure for detecting cancer metastases

on giga-pixel pathology images. With multi-grained analysis, the proposed method can capture rich information from pathology images.

4. We proposed a novel attention mechanism, semantic attention (SeA) to separate the proposals for saliency object detection. The attention is established based on the observation that regions with high attention should have similarly semantic concepts with salient objects.
5. Extensive comparative experiments have been conducted and the results shown that the proposed attention methods achieved excellent performance

4.2 Future work

On the one hand, we will further explore the attention mechanism to improve the feature-aware deep learning model design for detection and segmentation tasks. Due to the lack of well-labelled training data in many application scenarios, using the attention mechanism to enhance the positioning of weakly supervised and semi-supervised learning will be one of our research directions. At present, the self-attention-based vision transformer (ViT) [54] has shown excellent performance in various computer vision tasks. ViT provides a simple solution to learn cross-attention correspondence and fuse features from different modalities by applying multi-head fully connected layers. In addition, we will continue to explore the attention mechanism, make full use of the low-level and high-level semantic features of the image, and improve the performance of the deep neural network.

On the other hand, all research is aimed at dedicated application scenarios. We will further explore AI-aided diagnosis solutions for pathological images. According to the review article [48], DL based pathology analytics for Breast Cancer are still focused on tumor detection, the basic image analysis and at internal validation stage. To enable the advanced clinical applications, we will explore more advanced segmentation frameworks and backbone networks, build an end-to-end learning framework to predict the pN stage of WSI and achieve accurate classification and subtype clinical applications. We will introduce weakly supervised learning and semi-supervised learning based on the attention mechanism into the model to improve performance.

Acknowledgement

First of all, I would like to express my sincere appreciation to my supervisors, Associate Professor Tian Song and Professor Takashi Shimamoto, who have continuously given me valuable guidance, inspiration, advice, and criticism throughout my research work. Without their encouragement and support, I would not have been able to complete this work.

I would also like to express my deep gratitude to Professor Masaki Hashizume, Professor Yoshifumi Nishio, Associate Professor Hiroyuki Yotsuyanagi, Associate Professor Yoko Uwate, Assistant Professor Takafumi Katayama, and other technical staffs in our course, for their supports and friendship.

Thanks are due to past and present members of our laboratory for their help and fellowship. Special thanks must be paid to Dr. Xiantao Jiang, and Dr. Wen Shi for their dependable partnership, technical guidance, and helpful suggestions.

I wish to thank the members of my dissertation committee, thanks them for taking time to attend this oral defence, their valuable and helpful comments will guide me in the future research work.

I would like to give my sincerely thanks to my family, my father, mother, husband and daughter, thanks for your always supporting in my study. Without you I couldn’t be able to gain these achievements. I wish you all a healthy body.

Finally, I would like to express my gratitude to all friends who support me.

Reference

[1] K. Chang, T. Liu, H. Chen, and S. Lai: Fusing generic objectness and visual saliency for salient object detection. in Proceedings of IEEE International Conference on Computer Vision, pp. 914–921, Barcelona, 2011.

[2] S. Li, H. Lu, Z. Lin, X. Shen, and B. Price: Adaptive metric learning for saliency detection. IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3321-31, 2015.

[3] P. Jiang, H. Ling, J. Yu, and J. Peng: Salient region detection by ufo: Unique- ness, focusness and objectness. in Proceedings of IEEE International Conference on Computer Vision, pp. 1976–1983, NSW, 2013.

[4] P. Kr¨ahenbu¨hl and V. Koltun: Geodesic object proposals. in Proceedings of European Conference on Computer Vision, pp. 725–739, Springer, Switzerland, 2014.

[5] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk: Frequency-tuned salient region detection. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604, FL, 2009.

[6] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu: Global contrast based salient region detection. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–416, CO, 2011.

[7] Z. Liu, W. Zou, and O.L. Meur: Saliency tree: A novel saliency detection framework. IEEE Transactions on Image Processing, vol. 23, no. 5, pp. 1937–1952, 2014.

[8] Q. Yan, L. Xu, J. Shi, and J. Jia: Hierarchical saliency detection. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155–1162, OR, 2013.

[9] Y. Wei, F. Wen, W. Zhu, and J. Sun: Geodesic saliency using background priors. in Proceedings of European Conference on Computer Vision, pp. 29–42, Springer, Heidelberg, 2012.

[10] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang: Saliency detection via graph-based manifold ranking. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173, OR, 2013.

[11] W. Zhu, S. Liang, Y. Wei, and J. Sun: Saliency optimization from robust background detection. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2814–2821, OH, 2014.

[12] Z. Jiang and S. Davis: Submodular salient region detection. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2043– 2050, OR, 2013.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton: Imagenet classification with deep convolutional neural networks. in Advances in Neural Information Processing Systems, pp. 1097–1105, NV, 2012.

[14] S. Ren, K. He, R. Girshick, and J. Sun: Faster r-cnn: Towards real-time object detection with region proposal networks. in Advances in Neural Information Processing Systems, pp. 91–99, Montreal, 2015.

[15] K. Simonyan and A. Zisserman: Very deep convolutional networks for large-scale image recognition, Computer Science arXiv:1409.1556, 2014.

[16] C. Rother, V. Kolmogorov, and A. Blake: Grabcut: interactive foreground extraction using iterated graph cuts. Acm Transactions on Graphics, vol. 23, no. 3, pp. 309-314, 2004.

[17] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum: Learning to detect a salient object. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 2, pp. 353–367, 2011.

[18] V. Movahedi and J. H. Elder: Design and perceptual validation of performance measures for salient object segmentation. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 49–56, CA, 2010.

[19] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A.L. Yuille: The secrets of salient object segmentation. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 280–287, OH, 2014.

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

[21] N. Tong, H. Lu, X. Ruan, and M. Yang: Salient object detection via bootstrap learning. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1884–1892, MA, 2015.

[22] Y. Qin, H. Lu, Y. Xu, and H. Wang: Saliency detection via cellular automata. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 110–119, MA, 2015.

[23] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li: Salient object detection: A discriminative regional feature integration approach. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2083–2090, OR, 2013.

[24] X. Li, H. Lu, L. Zhang, X. Ruan, and M. Yang: Saliency detection via dense and sparse reconstruction. in Proceedings of IEEE International Conference on Computer Vision, pp. 2976–2983, NSW, 2013.

[25] J. Kim, D. Han, Y. Tai, and J. Kim: Salient region detection via high- dimensional color transform. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, OH, 2014.

[26] L. Wang, H. Lu, X. Ruan, and M. Yang: Deep networks for saliency detection via local estimation and global search. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3183-3192, MA, 2015.

[27] C. Li, Y. Yuan, W. Cai, and Y. Xia: Robust saliency detection via regularized random walks ranking. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2710–2717, MA, 2015.

[28] T. Wang, L. Zhang, H. Lu, C., Springer, Sun, and J. Qi: Kernelized subspace ranking for saliency detection. in Proceedings of European Conference on Computer Vision (ECCV), pp. 450–466, Amsterdam, 2016.

[29] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu: Spatial transformer networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS) - Volume 2, MIT Press, pp. 2017 - 2025, MA, 2015.

[30] Sanghyun Woo , Jongchan Park , Joon-Young Lee and In So Kweon: CBAM: Convolutional Block Attention Module. Computer Vision - European Conference on Computer Vision, Volume 11211, 2018.

[31] J. Hu, S. Li, S Albanie, S. Gang, and E. H. Wu: Squeeze-and-Excitation Networks. in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[32] H. W. Sang, Q. H. Zhou, Y. Zhao: PCANet: Pyramid convolutional attention network for semantic segmentation, Image and Vision Computing, Volume 103, 2020.

[33] X. Zhu, D. Cheng, Z. Zhang, S. Lin and J. Dai: An Empirical Study of Spatial Attention Mechanisms in Deep Networks, IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 6687-6696, 2019.

[34] A. Vaswani, N.M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin: Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), pp 6000–6010, NY, 2017.

[35] R. M. Cong, J. J. Lei, H. Z. Fu, M. M. Cheng, W. S. Lin, Q. M. Huang: Review of visual saliency detection with comprehensive information, IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), 2018.

[36] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and

fully connected CRFs, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 834_848, 2017.

[37] L. C. Chen, G. Papandreou, F. Schroff, H. Adam: Rethinking Atrous Convolution for Semantic Image Segmentation, Computer Vision and Pattern Recognition (cs.CV), 2017.

[38] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia: Pyramid scene parsing network, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881-2890, 2017.

[39] H. Li, P. Xiong, J. An and L. X. Wang: Pyramid Attention Network for Semantic Segmentation[J]. arXiv preprint arXiv:1805.10180, 2018.

[40] J. Fu, J. Liu, H. Tian, Z. Fang, & H. Lu: Dual Attention Network for Scene Segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3141-3149. 2019.

[41] J. L. Fleiss and J. Cohen: The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability, Educational and Psychological Measurement, vol. 33, no. 3, pp. 613-619, 1973.

[42] B. Lee and K. Paeng: Breast cancer stage classification in histopathology images, in Submission Results Camelyon17 Challenge, Nov. 2019. [Online]. Available: https://grand-challenge-publicprod.s3.amazonaws.com/evaluation-supplementary/80/dd7ab09f-dcdf-4bf6-be6d-2b070c9bfbd4 /lunit_camelyon17_paper.pdf.

[43] K. He, X. Zhang, S. Ren, and J. Sun: Deep residual learning for image recognition, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770_778, 2016.

[44] O. Ronneberger, P. Fischer, and T. Brox: U-Net: Convolutional networks for biomedical image segmentation, in Proc. MICCAI in Lecture Notes in Computer Science, pp. 234_241, 2015.

[45] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang: UNet++: A nested U-net architecture for medical image segmentation, in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (Lecture Notes in Computer Science), vol. 11045, D. Stoyanov et al., Eds. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-00889-5_1.

[46] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu: A density-based algorithm for discovering clusters in large spatial databases with noise, in Proc. Trans. Knowl. Discovery Data, (ACM), pp. 226_231, 1996.

[47] S. Lee, S. Oh, K. Choi, and S. Kim: Automatic classification on patient level breast cancer metastases, in Submission Results Camelyon17 Challenge. Nov. 2019. [Online]. Available: https://grand-challengepublic-prod.s3.amazonaws.com/evaluation-supplementary/80/46fc579c-51f0-40c4-bd1a-7c28e8033f33/Camelyon17_.pdf.

[48] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather: Deep learning in cancer pathology: A new generation of clinical biomarkers, Brit. J. Cancer, vol. 124, no. 4, pp. 686–696, Feb. 2021.

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

[50] G. Li and Y. Yu: Deep Contrast Learning for Salient Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 478-487, 2016.

[51] N. Liu and J. Han: DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 678-686, 2016.

[52] T. Zhao and X. Wu: Pyramid Feature Attention Network for Saliency Detection, IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3080-3089, 2019.

[53] B. Xu, H. Liang, R. Liang, & P. Chen: Locate Globally, Segment Locally: A Progressive Architecture With Knowledge Review Network for Salient Object Detection. Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3004-3012, 2021.

[54] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, & N. Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv, abs/2010.11929*. 2021.

[55] https://github.com/dmlc/xgboost.

[56] H. Zhang and I. J. Goodfellow and D. N. Metaxas and A. Odena: Self-Attention Generative Adversarial Networks. Proceedings of the 36th International Conference on Machine Learning. pp. 7354-7363. 2019.

[57] X. Wang, R. Girshick, A. Gupta and K. He: Non-local Neural Networks, IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794-7803, 2018.

[58] N.K. Singh, & K. Raza: Medical Image Generation Using Generative Adversarial Networks: A Review. *Health Informatics*, 2021.

[59] T. Wang, Y. Lei, Y. Fu, J.F. Wynne, W.J. Curran, T. Liu and, X. Yang: A review on medical imaging synthesis using deep learning and its clinical applications. Journal of Applied Clinical Medical Physics (JACMP), pp. 11-36. 2021.

[60] A. Singh and S. Sengupta and V. Lakshminarayan: Explainable deep learning models in medical image analysis. Journal of Imaging, volume 6. 2020.

[61] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang: CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. IEEE transactions on medical imaging, 40(2), pp. 699–711. 2021.

[62] Y. Peng, K. Yan, V. Sandfort, R.M. Summers, and Z. Lu: A self-attention based deep learning method for lesion attribute detection from CT reports. In IEEE International Conference on Healthcare Informatics (ICHI), pp. 1-5, 2019.

[63] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.J., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., & Rueckert, D: Attention U-Net: Learning Where to Look for the Pancreas. ArXiv, abs/1804.03999, 2018.