Deep Learning Based Anatomical Structure Localization and Segmentation in Fetal Ultrasound Images

胎児の超音波画像における ディープラーニングに基づく 解剖学的構造検出と分割

March 2019

Waseda University

Graduate School of Creative Science and Engineering

Department of Modern Mechanical Engineering, Research on Image Engineering

Yan LI

Acknowledgments

I would like to express the deepest gratitude to my supervisor Professor Dr. Jun Ohya, who has given me advices to complete this research. He provides a good study environment and encouragement to the research directions during my Ph.D course at Department of Modern Mechanical Engineering, Waseda University. Besides the directly valuable helps to this research, his rigorous knowledge, enthusiasm for learning and personal integrity deserves my admiration in my life time.

I am also grateful to my doctoral thesis examination members: Honorary Professor Dr. Hiroshi Nagahashi of Tokyo Institute of Technology, Japan; Professor Dr. Shigeru Uesugi of Waseda University, Japan; Professor Dr. Hiroyasu Iwata of Waseda University, Japan. I appreciate their acceptance to be reviewers and helpful comments for the doctoral thesis reviews and examinations.

I would like to express special thanks to Professor Dr. Hiroyasu Iwata, who has given me big helps to the experiment data used in this research. I would also like to say special thanks to Professor Dr. Artus Krohn-Grimberghe of Paderborn University, Germany. He provides many valuable comments to the paper writing and research ideas. I am also highly indebted to Dr. Rong Xu, who has given me suggestions to the research direction.

In addition, many thanks to all the past and present members in Ohya Lab, and faculty members of Waseda University who helped me during the past years for the necessary supports.

Yan LI

March, 2019

Contents

List of Figures

List of Tables

Symbols

P	Position of uterus
I	Image sample
P_{Init}^{ij}	Initial position
P_{offset}^{ij}	Offset distance
X	Coordinate of x axis
Y	Coordinate of y axis
C^{ij}	Confidence of bonding box
$n.$	Number of reference boxes
P_{Pred}	Predicted position of uterus
$F()$	Deep learning model
V^{ij}	The determined value
θ	Preset threshold
W	Width of the input image
H	Height of the input image
S^{ij}	Area of bounding box
D	Output vector of network
L	Loss in network training
T	Ultrasound image dataset

IOU	Intersection of union
AP	Average precision
TPR	True positive rate
FPR	False positive rate
v	Pixel-wise label
x	Image pixel
Q	Number of pixels
p	Pxiel-wise confidence
N	Number of image samples
ω	Weight of true positive sample in loss function
M	Visualized discriminative map
w	Weight of network
f	Feature map
c	Predicted image-wise confidence
y	Ground truth label of input image

Summary

Among various medical devices, ultrasonoscopy is the most widely used imaging modality for antenatal examination, because it is harmless to human tissue and can obtain real-time results. The accuracy of ultrasound (US) examination relies on radiologist with years of experience. Lacks of professional training courses could cause to increase the risk of misdiagnose. Moreover, the manpower storage with clinical experience in hospitals is becoming serious issues for many countries all over the world. Thus, to provide more valuable examinations, the working efficiency of doctors needs to be improved by automatic systems. Such automatic system heavily relies on the performance of related medical image processing techniques, which still have gaps with human doctors. Therefore, this thesis aims at providing accurate semantic information of specified anatomical structures for robotic medical care, such as automatic antenatal examination system. To this end, conventional medical image processing methods have drawbacks on accuracy, running speed, etc. On the other hand, with the fast development of deep learning algorithms and hardware acceleration techniques, the computer vision area has achieved great success in recent years. However, deep learning techniques have not yet been fully utilized for medical imaging domain. To solve the above-mentioned issues, this research proposes new deep learning based approaches for analyzing antenatal US images automatically.

The proposals of this thesis are categorized in the following three aspects:

Location of uterus The location and the border of the uterus are important for subsequent processes such as the segmentation of anatomical structures, and the location based guidance of a US probe in automatic medical care systems. Challenges in localizing of uterus from US images include noises and irregular shapes of the target object. The noises may lead to blurred areas in the border of the uterus and incorrect appearances of tissue structures. The uterus has irregular shape because of the non-rigid tissues and different view angles.

Areas of amniotic fluid and fetal body The areas of amniotic fluid and fetal body provide

important physiological indexes which can reflect physiology changes in the fetus, and can be used as the guidance of the probe in antenatal examinations. Challenges of the segmentation of the anatomical structure include noises and artifacts in the border areas of the fluid and fetal body. In addition, similar appearance of the body tissue and other tissues such as uterine wall of pregnant women could cause adhesion in adjacent blobs.

Area of fetal head The region of the fetal head provides the fine grained information on the fetal face and brain. The shape and the appearance of the fetal head can be adopted to diagnose fetal hydrocephalus and/or brain tumor. The shape and the position of the fetal head are important for automatic fetal care systems. Such a system requires technologies that locate the fetal head so as to infer the gesture and position of the fetus. It is difficult to determine the classification hyper plane of the slices that include the fetal head, because the appearances of the most of the slices of fetal head are easy to be confused with other fetal body parts such as abdomen slice, etc.

Derived from the above-mentioned issues, the proposed deep learning based approaches target at filling the blank areas of related tasks in antenatal examinations and optimizing the existing deep learning methods for US image areas. The scheme is separated into the following three modules: 1) bounding box regression Convolutional Neural Network (CNN) for uterus localization, 2) segmentation CNN for semantic segmentation of multiple anatomical structures, and 3) weakly-supervised module for region mining of fetal head.

The proposed methods and the relationship among the three modules are explained below.

1) Bounding box regression CNN for the uterus localization:

The accurate position of the uterus can be used as the region of interest for the subsequent processes such as the semantic segmentation of the anatomical structures. It is difficult to learn shape information from non-rigid objects by handcraft feature descriptors. The existing deep learning based object detectors are mainly used for natural image areas and lack of alignment accuracy. This module proposes a novel method for accurately locating the bounding box of the pregnant uterus in US images.

The proposed deep learning based method utilizes off-the-shelf CNN architecture as the backbone network, and designs a specific regression output structure to regress the candidate positions of the uterus. In particular, to obtain the abundant positions information of the uterus in US image, multiple densely positioned reference boxes are assigned according to the original image. The output of the network is designed as a vector which has same length as the coordinates and confidence of all of the reference boxes. Note that, to enhance the global context information, the output vectors are obtained through linear combination with fully connected weights. During the training phase, the weights of the network are assigned to learn the offsets between pre-defined positions to the ground truth and the confidence of each of the reference box. During the testing phase, the multiple candidate positions of the uterus are regressed by using predicted offset vectors to transfer the pre-defined positions. As the post processing approach, the method seeks the final position by non-maximum suppression to eliminate the redundant candidates.

The proposed uterus localization method is verified using the pregnant US dataset which is collected from clinical examinations. Comparative experiments demonstrate higher detection accuracy than directly using the methods that are to be applied for natural images. Other than that, the method achieves better alignment to the uterus area.

2) Optimized framework for semantic segmentation of multiple anatomical structures:

It is difficult to adopt local feature or cluster analysis based approaches to achieve accurate pixel-wise segmentations in US images. Related deep learning based methods for natural images still have room to improve on segmentation accuracy and smoothness in US images. To provide more accurate and smooth location information of multiple anatomical structures such as the uterus, amniotic fluid and fetal body in pregnant US images, this thesis adopts a deep learning based semantic segmentation framework and proposes specifically designed optimizations.

The segmentation CNN first encodes the input US image into down-scaled feature maps; then adopts the symmetric designed up-scaling operations to map the feature maps back to the

original size to perform pixel-wise classifications. The final predicted masks of multiple anatomical structures are obtained by threshold on the confidence map of each of the categories. Through preliminary experiments on US images, the study finds that the method needs to be further improved for better segmentation accuracy and smooth border areas.

The optimizations for above-mentioned issues are carried out by various ways: 1. The additional inner layers which can enhance the global representations; 2. The usage of the bounding box of the uterus detection which can reduce the data imbalance issue in the pixel-wise classification tasks; and 3. The multiple intermediate supervision layers which can bring obvious improvements to the smoothness of the segmentation blob.

The effectiveness of the proposed approach is evaluated by several different metrics such as IOU (Intersection Over Union), and ROC (Receiver Operating Characteristic) curve of each category on clinical US dataset. Compared with other related deep learning based methods, the proposed method achieves smaller errors to the ground truths (which are manual annotated by doctors with years of experience). In addition, the visualized results demonstrate smoother segmentations by comparing with the baseline methods in US images. The results of this work can be used to accurate reconstructions of fetuses or guidance of an automatic US probe.

3) Weakly-supervised methods for region mining of fetal head:

This module aims at implicitly learning the region of the fetal head based on image level annotations. The existing deep learning based weakly-supervised approaches have defects such as inaccurate localization and incomplete segmentation because of the discriminative area of the used feature level cannot represent the integrated region of the object. Therefore, an optimized method for fetal head plane classification and region mining by learning from image level annotations is proposed. To obtain more complete fetal area than existing works, the study proposes an optimized method which extracts the multiple hieratical feature maps as the discriminative area of the fetal head. In particular, to deal with the issue of incomplete segmentations in US images, a multiple output structure with different feature levels is

designed. The proposed weakly-supervised module merges multi-scaled discriminative maps with different feature levels to get more complete salient areas. By means of the proposed multiple output structure, final results are optimized through the combination strategy of multiple discriminative maps.

Comparison experiments are conducted on manual labeled fetal head US slices. Experiments demonstrate the method achieves high classification results and overlapping accuracy. Furthermore, the completeness of the obtained fetal head region is better than conventional related methods.

In summary, first, this thesis has introduced a deep learning based framework for fully supervised uterus detection in US images. Second, the experiments verify the effectiveness of various deep learning based methods for multi-category anatomical structure segmentation and has proposed optimizations for pregnant US images. Furthermore, this thesis has optimized weakly-supervised region mining of fetal head by merging multiple discriminative areas. To verify the gaps between the proposed methods and real-world usages, the performance of each module is compared with human doctors with years of experience. The results are promising that this research makes the development of automatic antennal examinations one step closer to the real world solutions.

The overviews of each chapter are listed as follows.

Chapter 1 describes the background and purpose of the thesis. And it makes brief introductions to the related works of the study, existing issues, and proposed technique modules.

Chapter 2 describes the related works from several aspects: the detection and semantic segmentation work in the medical image processing area, the development of related deep learning techniques applied in other computer vision areas, and the usage of deep learning techniques in the medical image processing area.

Chapter 3 explains the uterus detection module and demonstrates the results of experiments

for algorithm verification. The CNN structure for uterus detection and detailed training parameters can be found in this chapter, and, the models that trained under different settings are compared through experimenting on a pregnant US dataset.

Chapter 4 explains the proposed semantic segmentation of anatomical structure module. The algorithm is verified through a serial of experiments that conducted on clinical US dataset.

Chapter 5 proposes an optimized weakly-supervised region mining method for fetal head area discovering and localization. The preliminary experiments verified the effectiveness of the proposed method.

Chapter 6 summarizes the thesis and prospects the future development of the research.

Chapter 1. Introduction

1.1 Background

The antenatal examination is critical to pregnant women and fetuses. Through antenatal examinations, doctors can check growth and health condition such as the height and weight of the fetus [1] or the obstetrical complications of expectant mothers at various stages of pregnancies [2]. The antenatal examination is carefully scheduled at set intervals during the entire pregnancy. Lack of standard in the examinations will cause to increase the risk of misdiagnosing [3]. Therefore, physicians cost lots of efforts and times to inspection and analyze to the examination results. More recently, with increased health awareness of people from developing countries such as China and India, the shortage of manpower with clinical experience in hospitals is becoming big issues to many countries all over the world. The total amount of quality service resources for maternal and child health care is insufficient. To alleviate the situation, the working efficiency of doctors needs to be increased.

On the other hand, with the rapid development of medical imaging devices, computer diagnostic system and clinical diagnosis could be proceeded by various imaging modalities. Human doctors or computer-aided medical care systems rely on visual inputs as part of the reference to make diagnostic. Visualized human organs and tissues bring intuitive diagnostic references for doctors, and non-contact diagnostic applications bring less harm to patients. Many kinds of imaging modalities such as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and X-ray can be applied to medical examinations. Among various modalities, ultrasonography is commonly used, because it provides a real-time and intuitive reference to the doctors from the department of gynecology and obstetrics. What is more important, in antenatal care, the fetus and pregnant woman are

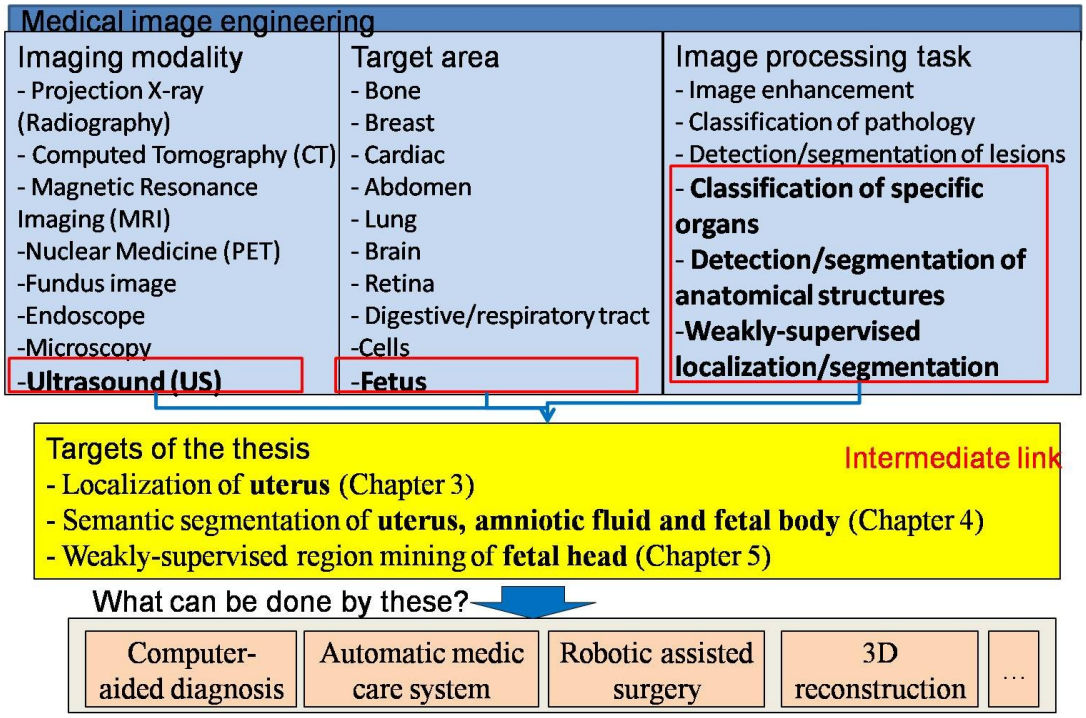

Figure 1.1 Position of the proposed targets and solutions in medical image processing area.

very sensitive to physical or chemical damages. The patient needs more medical care than in normal situations. Under this situation, the radiation based fluoroscopy image systems such as X-ray or PET-CT should be avoided as much as possible [4]. In contrast, the US device is non-invasive and harmless to both of the patient and fetus. And the US examination has advantages in relatively low cost, compared with other medical imaging devices such as MRI. Thus, most clinical antenatal examinations are preferred to use the ultrasound (US) imaging devices [5].

The fully automatic or computer-aided medical care system based on medical images can be expected to raise the efficiency of doctors' jobs. To achieve this target, image processing modules are critical to such a system. Therefore, in this research, the author proposes to adopt several high level semantic extraction modules, which could be used as the important intermediate links between upper level automatic medical care system and raw US image data, as shown in Figure 1.1.

In the early time, the research in medical image processing areas rely on low-level feature descriptors such as gradient and corner, as well as handcraft descriptors, in order to provide useful information for computer-aided systems or doctors. Those methods, however, have limitations for tasks that need to deal with more complex hidden features or higher precision requirements. In contrast, in recent years, the deep learning based approaches outperform the handcraft feature descriptors and traditional classifiers in many computer vision areas. The exploration of the usage of deep learning based technology in medical imaging becomes a concerning topic as well.

As a modern machine learning scheme, deep learning is first named by Geoff Hinton at the University of Toronto in 2006 [6]. To the image processing area, it is a starting point in which the deep learning technique has made considerable achievements in many tasks. The mechanism of share-able weights of the convolution neural network (CNN) reduces tons of learning space. And, the development of graphics processing unit (GPU) accelerates the computation by hundreds of times than traditional processing speed. The usages of these techniques and hardware have significantly contributed to the progress of most of computer vision tasks in the world, such as object recognition task [7][8][9], object detection task [10][11][12], and semantic segmentation task [13][14]. With the development of deep learning techniques, the study to the medical image processing is also being widely and successfully researched during the last few years. The performances of deep learning based algorithms have outperformed many of conventional approaches with large gaps. For example, Rajpurkar et al. [15] report their deep learning based algorithm achieves higher accuracy than expert teams that classify chest pathologies from chest x-ray images.

Inspired by related deep learning works, in this research, the author proposes a series of deep learning based medical image processing approaches to provide high level semantic features such as the location or the fine-grained area of desired anatomical structures in US images for both human doctors and computer-aided systems conducting antenatal examinations. In particular, from course to fine, this study focuses on the following three aspects: the localization of the uterus, the semantic segmentation of the amniotic fluid and fetal body, and

the weakly supervised region mining for the fetal head area. The medical indicators, appearance feature, challenges and proposed solutions of each related anatomical structure is introduced as follows:

Uterus In the prenatal care, the uterus is the most important anatomical structure for pregnancy tests and fetal examinations. As the major reproductive system, the uterus carries and protects the fetus. The uterus is supported by multiple tissues such as a pelvic diaphragm, perineal body, the peritoneal ligament and the broad ligament of uterus [16]. In US image processing, the location and the border of the uterus are important for the subsequent processes such as the segmentation of anatomical structures, for providing the location-based guidance with the US probe for automatic medical cares, and for helping to reveal the intrauterine disease for human doctors. As an instance, in the machine-aided amniocentesis (also referred to as amniotic fluid test or AFT) [17], in order to collect cells from the fluid and fetal tissue from the living body, the doctor lets a sampling probe go through the uterus from outside of the abdomen of pregnant subject. The sampling probe needs to pierce the abdominal cavity and uterine wall and to collect amniotic fluid from the uterus. The improperly implemented amniocentesis could cause amniotic fluid embolism precipitated [18]. To such systems, the size and position of the uterus can provide the most fundamental location information to guide the multiaxial servo controlled manipulator.

The uterus has an inverted triangle shape and is centered at the pelvic cavity [19]. For a healthy pregnant woman, the uterus includes several anatomical structures such as the uterine wall, amniotic cavity, and fetal body. The uterine wall is a tissue composed of soft structures. In US images, the appearance of pregnant uteruses presents an irregular form in a closed shape. The irregular shape of the appearance is caused by the different size of the fetus and the position in the abdominal cavity. Inside this closed area of the pregnant woman, the tissues of the fetus can be viewed as similar appearances to other body tissues. Dark areas in the uterus are amniotic fluid. The tissues and amniotic fluid have different tissue densities, which results in different appearances in US images.

The challenge in the localization of the uterus from US images includes image noises and irregular shapes of the target. First, the noises and artifact inevitably appear in medical US images because of various factors in imaging mechanisms such as the coherence property and reflections [20]. The noise and artifact may lead to blur the border of the uterus and to yield incorrect appearances of located tissue structures. Second, as above-mentioned, the uterus shows an irregular shape because of the soft structure of the tissue and different view angles. In general, it is difficult to describe and learn shape information on non-rigid objects.

In summary, in order to provide the base location information of uterus for subsequence approaches or automatic medical care system, the research proposes a specific module to localize the position of the uterus from the raw US images.

Amniotic fluid and the fetal body In addition to the location of the uterus, the automatic system desires to make further use of more high level information from the pregnant US images such as the shape and the position of amniotic fluid and fetal body. The amniotic fluid and fetal body, which are wrapped in the uterus, are important physiological indexes which can reflect physiology changes in pregnant women and fetuses. In particular, the amniotic fluid volume is the material that is filled in the uterine wall. The main component of amniotic fluid is alkaline liquid which consists of 90% of water and 10% of other materials such as urea, uric acid, and epithelial cells of fetus [21]. The number of specific proteins in amniotic fluid can be used as a marker for monitoring abnormalities of the fetus, and the estimation of the volume of the liquid can be used for measuring placental functions and to prevent body disorders such as hypolimnion and hydramnios [22]. Indicators for fetal body parts can be used for estimating the weight of the fetus and prediagnosing the fetal abnormalities. On the other side, in the amniotic fluid testing system, besides the position and shape of the uterus, the critical information is obtained from the accurate segmentation of the amniotic fluid and fetal body. The needle of the probe is required to be located inside the uterine wall and collects different samples from the amniotic fluid and part of the body tissue of the fetus. Therefore, the segmentation of the amniotic fluid and the fetal body in the uterus area in the US image is essential to the automatic medical care system.

Regarding the structure of different tissues in the uterus, all of the tissues of the fetus are soaked in the amniotic fluid. Most of the anatomical structures of the fetus can also be observed in the US slices. The shape and size of such anatomical structures in the US image are irregular because of the variation in view angles and fetal postures. Regarding the amniotic fluid, the amniotic fluid, which is dense liquid, is normally observed as dark areas inside the uterine wall. The fetal body tend to have complex shapes and appearances, because the fetus is bent inside the uterus; therefore, multiple tissues can be viewed in each slice, and the appearance and acoustic shadows of different tissues influence each other in US images.

The challenges of segmenting the anatomical structure in the US image are caused by several aspects. First, defects of imaging mechanisms cause noises and artifact in the border area of the uterus. The segmented blobs have unsmoothed borders. Then, similar appearances of the body tissue and other tissues such as the uterine wall of pregnant women cause the adhesion of the segmented blobs. As it is known, the terminal system of the US probe can provide 3D fetal images by removing the amniotic fluid. The system can judge the categories of materials by different densities of objects. However, even if the system can distinguish the amniotic fluid from other tissues, it is hard to separate the fetal body and uterus because the densities of those structures are similar.

Therefore, as one of the modules of the automatic medical care system, this thesis proposes and verifies a visual based semantic segmentation module to segment the specified structures in the US images of pregnant women.

Fetal head Among fetal body parts, the fetal head can be used for estimating the health status of the pregnant woman and fetus. Doctors adopt the size and shape of the fetal head to make a diagnosis for the patient and judge if the parturition will go well. For example, the biparietal diameter (BPD) is used as one of the parameters to estimate the weights and growth of the fetus [23]. The shape and appearance of the fetal head can be utilized to diagnose the hydrocephalus or brain tumor for the fetus. In addition, the relative position of the fetal head and pelvis is an important reference to the doctor to judge the appropriate way of parturition.

The shape and position of the fetal head are also important for automatic fetal care systems. Such a system requires a technology that locates the fetal head so as to infer the gesture and position of the fetus; then, the system can perform subsequent processes such as guiding the US probe to the specified positions for further measurements. In the above-mentioned amniotic fluid testing system, the needle needs to reach the position accurately inside the uterine wall and collects the amniotic fluid and fetal tissues. During the examination, the head of the fetus must not be touched by the probe because it is dangerous for the fetus. Therefore, the accurate identification of the fetal head area in the raw US image can help to guide the position of the needle to avoid hurting the important head tissues of the fetus.

The head of the normal fetus is bilaterally symmetrical and can be observed as a closed ellipse shape in the US image. The appearance of the fetal head temporally changes. Under normal condition, the complete hyperechoic of the skull and internal tissue structures (such as the midline structure and the thalamus) can be seen after 11 weeks [24]. In order to prompt the slice with the fetal head and highlight the entire head structure in the input image sequence for automatic examination systems, in this research, the target location of the object includes the entire region of the fetal head structure. This task is more difficult than only extracting the standard plane of the fetal head because the appearance of the fetal head is very easy to be confused with other fetal body parts such as abdomen slice by common classification models.

Regarding the challenges of the proposed work, except the difficulty above-mentioned, such as the image noises and artifact of US imaging devices, the variation of the viewpoint and the shadows occluding other tissues, the fetal head has extra difficulty in making annotations. The elliptical outline of the fetal head has similar appearance feature with other tissues such as the fetal abdomen, and the fetal head slices can be viewed by doctors only in part of the video sequence. The professions need more meticulous annotation and take more time to judge whether the input image includes a fetal head or not.

The annotation of the fetal head slice can save much more time and reduce the human cost for annotation works. Compared with the tight bounding box or pixel-wise annotation of the fetal

head area in US images, the image level annotations such as the slice with or without a fetal head are much easier for annotation workers. What is more important, distinguishing the fetal head slice requires less practical experiences than pixel-wise annotations. Therefore, this thesis proposes a weakly-supervised module to learn fine-grained annotations of the fetal head area (such as bounding box and pixel-wise classification) only by coarse annotations (image level labels).

1.2 Related work

The recognition of the category and position of the anatomical structure is very important for automatic surgery and computer-aided medical care systems. Such fine-grained location information can be used to aid the automatic surgery system or human doctors to conduct precise operations. The detection and segmentation of the specified structure in the medical image has been widely researched for a long time. With the fast development of the CNN and hardware devices, the deep learning technique has made considerable achievement in many of the computer vision areas. More recently, the deep learning based method also affects the development in medical image processing areas and arouse lots of attention to the researchers. More details of related convention and deep learning based works can be found in Chapter 2.

1.3 Motivation

As above-mentioned, the shortage of medical doctors who can conduct medical examinations has become a serious problem for hospitals all over the world. The targets of this thesis include improving the efficiency of doctors who conduct medical image based antenatal examinations and providing high-level semantic information (such as the location of specific anatomical structures and the category of the input US slice) for automatic medical care systems. More specifically, the purposes of each module that is proposed in this study are listed below.

1) The position of the uterus is required by automatic surgeries or computer-aided systems in antenatal examinations, and can be used as the region of interest for the subsequent processes such as the semantic segmentation of the anatomical structure. Existing works

(e.g. [29][30][31]) try to estimate the position of tissues or lesions in different ways. However, these existing methods are not robust enough to deal with objects such as the uterus, because it is hard to discriminate the irregular shape by manually designed feature descriptors. This thesis proposes a deep learning based method for locating the position of the pregnant uterus in US images. The bounding box of the uterus is localized through a regression-based CNN framework.

2) To provide the fine-grained location information for different anatomical structures in the uterus, this thesis proposes a semantic segmentation method for distinguishing the areas of the uterus, amniotic fluid, and fetal body. The related works [34][41] cannot achieve acceptable accuracy, because the local feature extracted from the US image is easy to be confused, and shape based methods [43][45] heavily rely on the initial positions. On the other hand, the precise segmentation information is important for automatic examination systems such as computer-aided AFT. Therefore, in order to improve the smoothness of the segmentation results and pixel-wise classification accuracy, this thesis further proposes an optimized approach for the accurate segmentation of specified anatomical structures.

3) Some of the existing methods use fully supervised approach to detect an object; however, the workloads of annotating fetal head areas cost large human powers. Thus, after 1) detecting the uterus and 2) segmenting different anatomical structures in the US image, the fine-grained classification, and region mining of the fetal head using image level annotations are performed as the third module. In preliminary weakly-supervised works [15] [54] in the medical image area, the performance of the works still needs to be improved, because the existing weakly-supervised approaches lack completeness of segmentation results. Therefore, to simplify the annotation and save work time, and to provide more feasibility for accurate localization by the weakly-supervised method, this thesis proposes an improved method for learning to mine the region of the target only using image level labels.

1.4 Introduction of Proposed Approach

To achieve the targets that described above and solve the related issues, this thesis proposes an approach that consists of three parts as follows.

1) Detection of uterus

The proposed deep learning based method utilizes a backbone CNN network and regression output structure to regress candidate positions of the uterus. Then, as the post-processing approach, the method seeks the final position by eliminating redundant candidates. The method adopts densely designed reference boxes to achieve abundant position information from US image to obtain the accurate uterus localization results for the subsequence modules.

- Uterus detection network

The proposed method utilizes advantage of the end-to-end learning task. It is designed to use the resized US image as the input and directly map the output into a vector with the fixed number of elements that correspond to the offsets of multiple pre-defined reference positions. Regarding the backbone network, the proposed method utilizes CNN with fully connecting output layer as the feature extractor. The output structure is composed of the weights which can map the feature vector to the same length vector containing the coordinates of the offsets and confidence scores. The offsets correspond to the top left and bottom right distances between the pre-defined positions and ground truth bounding boxes, and the confidence scores correspond to the probability that the uterus is at each position.

- Post-processing

The proposed method determines the candidate positions of the uterus by thresholding the output confidence scores. The accurate positions of each candidate are obtained by reshaped offset vectors and transferring operations. The final position of the bounding box is determined by eliminating the redundant boxes based on box overlapping and confidence score of each candidate.

2) Semantic segmentation of amniotic fluid and fetal body

The proposed deep learning based segmentation CNN is separated into two strategies (a binary category for the uterus as the preliminary research, and multi-category for amniotic fluid and fetal body). This thesis also proposes optimization schemes to improve the pixel-wise classification results and provide more smooth segmentations.

- Uterus segmentation network (Binary category)

As the preliminary research, the author follows the existing method applied for nature image area and explores the limitation of the methods. An encoding-decoding architecture is adopted to segment the uterus from US images. The uterus segmentation CNN first encodes the input US image into down-scaled feature maps, and then adopts symmetric designed up-scaling operations to scale the feature maps back to the original size to perform pixel-wise classification of the uterus. The weights of the network are learned from fully supervised annotations of the uterus area and use the per-pixel sigmoid and binary loss as the target function. The confidence map of the uterus is mapped by stacked convolution operations and probability output functions. The final binary mask of the uterus is obtained by thresholding the confidence map. Through experiments, several issues are discovered and discussed.

- Amniotic fluid and fetal body segmentation network (multi-category)

The main framework of the multi-category segmentation CNN also follows an encoder-decoder structure to perform pixel-wise classification in the input US image. In order to achieve the multi-category segmentation, the cost function of the framework utilizes the multinomial loss. The confidence maps of each of the category are calculated from the output of the model. The segmentation masks of each specified structure are obtained by maximum operations at each pixel. To solve the issues discovered in the preliminary research, optimizations are proposed by the author.

- Optimizations

This thesis points out that, obtaining the smoothed border and accurate segmentation is important for the subsequence processes; thereby, this thesis further embeds optimization

schemes into the segmentation CNN applied to anatomical structures in US images. Specifically, the following three optimizations are introduced due to respective reason: 1. additional inner layers which can enhance global representations; 2. the usage of bounding boxes for uterus detection, which can relieve the data imbalance issue in pixel-wise classification task; and 3. multiple intermediate supervision layers, which can optimize the smoothness of the segmented blobs.

3) Weakly-supervised region mining of fetal head

After obtaining the bounding box of the uterus and the result of the semantic segmentation of amniotic fluid and fetal body, this thesis searches for region information of the fetal head by a fine-grained localization task. Due to the complexity and time cost of the pixel-wise annotation works, this thesis proposes a weakly supervised learning method for region mining of the fetal head area using image level annotations.

- Classification of the fetal head slice

The classification of the fetal head is conducted by a basic classification CNN structure so that the result of classifying the fetal head is obtained by learning using image level annotations. In order to achieve the target of the fetal head region mining, some parts of the network structure are modified to make use of the learned feature maps to infer the region of the object.

- Region mining of the fetal head area using image level labels

The method for the region mining of the fetal head area utilizes the learned classification model. This classification model discards the fully connecting layers and replaces with one global pooling operation. The regions of the positive object are extracted from the cumulated responses of multiple stacked convolution layers. The learned parameters of the output layer are used as the weights for calculating the final cumulative map. This thesis proposes a CNN network with new sub-structures to solve the issue of incomplete segmentations. In addition, the final results are optimized through the proposed combination strategy of multiple

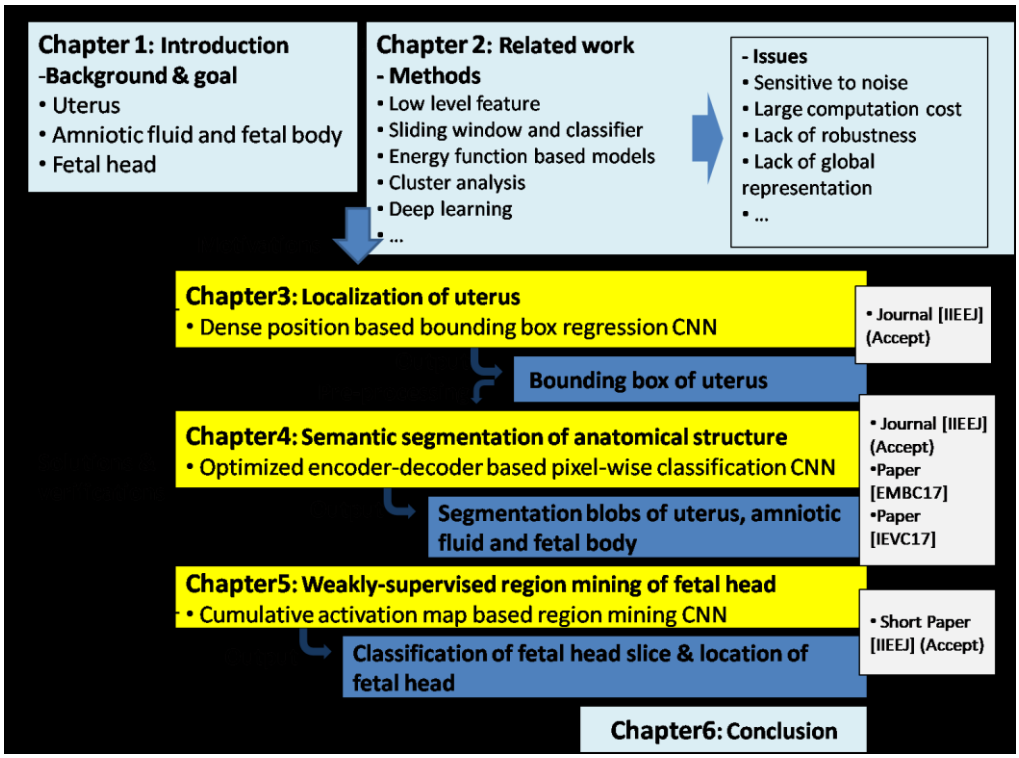

Figure 1.2 Organization of the thesis**.**

discriminative maps.

1.5 Organization of this Thesis

In order to provide important semantic information of target objects such as the location of the uterus and pixel-wise classification of anatomical structures in US images for automatic medical care systems, and to solve the related issues, this author proposes the three main modules in each of the following chapters, respectively, as follows. Chapter 3: detection of the uterus, Chapter 4: semantic segmentation of anatomical structures, and Chapter 5: weakly-supervised region mining of the fetal head. The relationship between the three modules and the rest of the chapters are illustrated in Figure.1.2.

Regarding the relationships of the proposed modules, first, all of the anatomical structures are located inside the uterus. The localization of the uterus is proposed as the first module. The localized area can be used as a coarse position for the subsequent semantic segmentation module. Therefore, the proposed method of localizing the position of the uterus in the raw US

image can be used as the pre-process for the segmentation of anatomical structures. Then, in order to further provide a more important area of the fetal head from the segmented fetal body inside the uterus, this thesis proposes the third module to localize the fetal head area by learning using image level annotations. Note that the proposed weakly supervised method can be extended to future research works and bring advantages to related areas.

An overview of each chapter, together with the publications, is described as follows.

Chapter 1 describes the background and purpose of this thesis. Then, a brief introduction to the related work of this thesis, existing issues, and the proposed three modules are described.

Chapter 2 describes related works in terms of several aspects: detection and semantic segmentation work in the medical image processing area, the development of related deep learning techniques applied in other computer vision areas, and the usage of deep learning techniques in the medical image processing area.

Chapter 3 explains the proposed uterus detection module and demonstrates the results of experiments for verifying the validity of the algorithm. The CNN structure for uterus detection and detailed training parameters are explained, and models that are trained under different settings are compared through experiments using a pregnant US dataset. As shown in Fig. 1.2, the contents of Chapter 3 are published in the journal [62] (to appear).

Chapter 4 explains the proposed module for the semantic segmentation of anatomical structures. The algorithm is verified through experiments conducted using a clinical US dataset. The part of binary segmentation of uterus area is published in [63], and the segmentation of multiple anatomical structures and its optimization methods are published in [64] and [62] (to appear).

Chapter 5 proposes an optimized weakly-supervised region mining method for finding and localizing the fetal head area. The preliminary experiments verified the effectiveness of the proposed method. The contents of this chapter are published in [65] (to appear).

Chapter 6 summarizes this thesis and prospects the future development of the research.

Chapter 2. Related Work

The related works described in this chapter are organized as follows: the traditional method based object detection and segmentation work in US images are introduced in Section 2.1 and 2.2. The related deep learning works applied by other computer vision areas are introduced in detail in Section 2.3. Then, the last section lists state-of-the-arts of deep learning based medical image processing works related to author's research.

2.1 Detection of anatomical structure in US image

The object detection technique can be used to assist human doctors or computer-aided medical care systems by providing useful location or region information of specific anatomical structure or lesions from raw US image data. The task is closely related to locate the spatial information of the object with a specific appearance that different from other background areas. The introductions of related works are as follows.

- Template matching

The template matching is widely used to detect target object in US image [26]. The S. Yu et al [25] use template matching based method to achieve lumber detection approach. The authors of [66] propose an adaptive function to detect the boundary of the brachial artery in US scans. The authors improve the traditional method by the adaptive designed template to suit the changeable shape of the target object, but the accuracy of the template matching approaches is still very sensitive to the position and relies on rigid shape information of the target object. The authors of [27] extend the template matching to a 3D based approach to detect the boundary of a kidney from sequential captured kidney US images. The method adopts the method named as PKSM to estimate the initial position of the kidney. The PKSM method fits

3D voxels of kidney's statistic template to the position which has the maximum probability in the volume of US data. The model can locate the 3D position in volume data by template matching. However, the method uses related small-scale training data, and it is only feasible for the simple and independent structures such as the kidney. In their proposal, it is also hard to determine the kidney exists or not.

- Object classifier and sliding window

The sliding window method divides the input image into multiple sub-regions with overlapped areas and performs object classification in each of the windows. For example, in 2015, N. B. Albayrak et al. et al proposed to use handcraft feature descriptor (HOG) to extract the responses of detection widows at each of the positions. The algorithm predicts the probability at each of the positions using a support vector machine (SVM) classifier to obtain the heat map of the anatomical regions in US images [67]. In another similar work, B. Rahmatullah et al. use local haar-like [30] feature descriptor and AdaBoost [68] as a classifier to perform a sliding window based fetal stomach and umbilical vein detection from a cross-section of fetal US slices [28]. The harr-like feature can achieve relatively fast feature extraction speed, while it is easy to have many false positives because of the lack of discriminative representations.

Another representative work by G. Pons et al. [29] propose to use a method with more advanced and complex approaches than using single object classification based method. In their work, Pons et al. utilize the deformable part based model (DPM [69]) to detect the position of the lesion. The DPM provides the detection score of part of the object and uses chain model to learn the adaptive model between different components. However, the bottleneck of the DPM based approaches is the redundant computation on a large number of windows, which causes the methods are hard to be transferred to real-world solutions.

Basically, these methods follow the object detection schemes that are used in nature images. In order to ensure the target object can be contained in the windows, the scales and the positions of the predefined windows are critical to the performance of related approaches. The

oversampled sliding window causes the redundant computation, while the insufficient sliding windows cannot estimate accurate alignment of the detection because the windows are too sparse for the location of the ground truth position.

- Region proposal

To deal with the issues of redundant computation for sliding window based approaches, some researchers choose to start their detection works from the pre-generated region of interest (ROI). R. Bharath et al. [31] propose a corner point descriptor based method to extract the region of interest for detecting fetal genital organs. W. Mahmud et al. [32] also propose to use automatic generated ROI to speed up and constrain the searching region to improve the accuracy of the detection of the kidney in US images. In Mahmud et al.'s work, the one or multiple seed regions are extracted by binary image based region mining. The candidate windows are extracted by prior knowledge of the appearance of the kidney in US images. The W. Mahmud et al.'s proposed method uses many handcraft parameters and prior knowledge; thereby, it lacks generalization to other anatomical structures.

The classification results in each of the window are affected by the local feature of the object because the feature of the medical image is different from the nature image because of the nonrigid appearance of the anatomical structures. In addition, another demerit of the methods is the heavy dependence on the ROI extractor. The recall of the detection results is limited by the quality of the proposed candidate windows.

2.2 Fully automatic segmentation of anatomical structure in the US image

The segmentation can be seen as the pixel-wise classification task. Compared with object detection task, fine-grained shape or area information can provide more valuable information to doctors or automatic systems. This thesis addresses the developments and issues in fully automatic segmentation applications without any manual interventions. The related works are detailed as follows.

- Region growing and watershed

In order to locate the common area belonging to the same object, the region growing method relies on the seed position to generate multiple small groups of pixels as regions and combines multiple independent regions into larger one. In particular, the starting positions of the region growing methods are based on the seed regions; then, the pixels or areas with the same priority are gradually merged into the same blob. P. R. Thangaraj et al. propose a watershed-based method to use seed regions for identifying and classifying the areas of renal calculi [34]. The same authors further optimize the watershed by ANFIS [70] method in [35] and [36]. Similarly, P. T. Akkasalgar et al. [71] utilize the region growing method to segment the area of a kidney in US images and use it as the pre-processing of the detection of the abnormal objects. Although sometimes with no prior knowledge can be used, the advantage of region growing based method is that the good performance still can be obtained. However, to the region growing algorithm, the computation cost is relatively large, and the speckle noises and gray level heterogeneity may lead to over-segmentation.

- The low-level feature descriptor

Low-level features such as edge, corner, or gradient can be used as good prior for segmenting the specific area from the image. For example, [39] proposes to use Canny [40] as a feature extractor to extract the contour of the follicles from US images. Although the low-level feature based methods or morphology operations such as Canny operator or region growing method have advantages in label-free mechanism, however, the local appearance based feature is very easy to be affected by speckle noises or blurred borders in the US image, because the manually designed feature descriptors are very sensitive to the pixel-wise variance and lack of prior knowledge of global shape.

- Cluster analysis

Similar to region growing based methods, cluster analysis is an unsupervised segmentation method. People do not need to manually add labels to each sample in the dataset. The cluster structure is automatically discovered by discriminative feature. N. Archip et al. [72] make use of spectral clustering based normalized cut (NuCut) method to segment fetus and abdominal

on simulated US image. Recently, in 2011, related optimizations are published in [73]. The method first splits the raw image into multiple continuous subregions, then uses curvelet transform and GLCM to extract various feature vectors, the spectral cluster is performed based on the feature, at last, they use KNN to segment the pathological area in US images. The spectral cluster has advantages than the common k-means method in high adaptability to data distribution and lower computation cost. J. Shan et al. adopt the optimized feature extractor and neutrosophic cluster to segment the lesions in breast US images in their works [41][42] and [74].

The merit of clustering based methods is its un-supervised learning scheme brings relatively fast training speed, while the method is sensitive to the isolated cluster and lack of evidence for choosing the suitable number of clusters and initialization of cluster centers.

- Energy function based contour extraction

The energy function based method refers to the active contour model [75] and its derivations. The basic idea is to use a continuous curve with multiple landmarks to express the edge of the target and define energy function to make the variable shape includes the curve of the object. In particular, the active contour model relies on multiple landmarks to extract the feature from the gradient with a high variance that fits the statistic model of the target object. The method is widely used in human face related image processing tasks such as feature point extraction. In medical image area, G. Slabaugh et al. make use of energy function based statistic model for extracting the contour of specific tissues from US images [43]. In addition, G. Slabaugh et al. further optimize their active contour-based method in [44]. In their proposed method they use active contour model based on whiten images and fisher-tippett distribution feature. The highlight of the research is the noise reduction pre-processing approach. However, the feature is still hard to deal with the structures with an irregular shape.

The active shape model is suitable to fit the shape of the object with closed contours and regular shape. The defect of the active contour model-based method is the heavy dependence on the initial positions and diversity of statistical models.

Table 2.1 Summarize of related works

Tasks	Methods	Authors	Issues
Detection of anatomical structure	Template matching	S.Yu et al, [25] L. Fan L et al, [66] M. Marsousi et al, [27]	-only feasible to the simple and independent structures
	Object classifier and sliding window	N. B. Albayrak et al, [67] B. Rahmatullah et al, [30] G. Pons, [29]	-redundant computation -cannot estimate accurate alignment
	Region proposal	R. Bharath et al [31] W. M. H. W. Mahmud et al [32]	-heavily dependence of the ROI
Fully automatic segmentation of anatomical structure	Region growing and watershed	P. R. Thangaraj et al, [34][35][36] P. T. Akkasalgar et al, [71]	-Not robust to the spackle noise and gray level heterogeneity
	Low level feature descriptor	P. S. Hiremath P S et al, [39]	-not robust to speckle noise and blurred borders -lack prior knowledge of global shape
	Cluster analysis	N. Archip et al, [72] H. –D. Cheng et al, [41][42]	-sensitive to the isolate cluster -lack of evidence for choosing the suitable number of clusters and initialization of cluster centers.
	Energy function based contour extraction	G. Slabaugh et al, [43][44]	- hard to deal with the structures with irregular shape - heavily affected by initial position

The above introduced related works are summarized in Table 2.1.

2.3 The development of related deep learning techniques applied to other computer vision areas

As mentioned above, deep learning based techniques achieve great success in many tasks of

computer vision area. This thesis tries to explore the feasibility of the related deep learning approaches of proposed targets. As a basic survey in this research, the development and common usage of deep learning methods for other computer vision areas are introduced in this section.

Backbone network and object classification Real-world solutions with deep learning based model are started to be noticed by researchers early in 1998. Y. L. Lecun et al. propose a neural network structure with stacked convolutional operations in their publication [76]. The proposed deep learning model is adapted to recognize a large amount of hand-written digits and demonstrates high classification accuracy. Compared with traditional neural network approaches, it utilizes convolutional operations with shared kernels to reduce the volume of learn-able parameters to a relatively acceptable range. After that, in 2012, A. Krizhevsky et al. [7] used a more complex CNN structure to classify a common object in nature images and achieve the best score in Imagenet [47] challenge. In their work, they propose a deeper network structure with more convolutional layers with learn-able weights and efficient training method by making use of multiple GPU devices. After that, in the following works that relate to the backbone network, the structure becomes more complex, such as VGG [48], googleNet [8], RESNet [9] etc. The networks are not only deeper than before, but also more efficient in learning from large-scale training data by optimized activation functions, loss functions and skip connection mechanisms etc. Researchers make use of end-to-end learning strategy to achieve successes in many of the computer vision tasks. Since then, the efficient training method, and related hardware devices have been widely researched and spreading to the field of computer vision in very fast speed.

Object detection Using CNN classification model on each position of the sliding window is a high computation cost choice. Compared with sliding window fashion, modern convolutional object detectors tend to detect objects in an end-to-end pipeline. The methods adopt a classification and regression network with shared convolution kernels to provide end-to-end learning scheme to regress the positions of foreground objects and classify the proposed candidates into specific categories. For example, the series of CNN detectors proposed by R.

Table 2.2 Summarize of related deep learning based approaches

Tasks	Methods	US images	Introduction
Object localization	Faster-RCNN [10]	×	- RPN based two-stage object detector.
	SSD [11]	×	- Fully convolution single-stage object detector.
	J. M. Wolterink et al, [83]	○	- Detect the coronary artery calcium in 3D US image volume by 3D convolution network
Semantic segmentation	Segnet [14]	×	- Symmetric designed encoding-decoding structure
	FCN [13]	×	Iterative trained fully convolution models
	DeeplabV3+ [95]	×	- Optimized by atrous convolution kernels
	PSPNet [96]	×	- Optimized by pyramid spatial pooling modules
	G. Carneiro G et al, [52]	○	- Deep belief network (DBN) based segmentation framework to extract the area of left ventricle
	H. Chen et al, [53]	○	- FCN based segmentation framework to extract the area of left ventricle
Weakly-supervised localization	S. Karen et al, [58]	×	- Visualize the discriminative area by back propagation
	M. Oquab et al, [57]	×	- Visualize the discriminative area by global max pooling
	CAM & its variants, [55] [56]	×	- Visualize the discriminative area by discriminative mapping
	Sononet & its variants, [54][59][60]	○	- Optimized back propagation [58] based method in US images
	N. Toussaint et al, [61]	○	- Directly adopt [56] in US images

Girshick et al. [78] [77] [10]. Their method treats the object detection problem as region proposal and object category classification phases. The position of the object is obtained

through prediction of end-to-end defined offsets between the ground truth and multiple anchors, and each anchor also predicts the confidence of the objectness. The final positions of objects are determined by further regressed offsets and confidence of object categories. There are other works such as [11] [12] that compress the detection pipeline into a single stage. In particular, similar to the first half stage of [10], in a single stage based convolutional object detectors, the position and category of the object are directly regressed and classified at the same time in jointly learned convolution layers. The final bounding boxes of the objects are obtained by transferring the multiple pre-defined positions (named as anchors) with regressed offsets. The methods show success in many object detection challenges.

Semantic segmentation Object segmentation can be seen as the fine-grained object detection task. The category of each of the pixels is required to be predicted. To obtain the best segmentation results without losing spatial information, the prediction results and original input image have a one-to-one correspondence in the pixel level. Similar to the classification or object detection tasks, the basic concept of deep learning based segmentation model also follows the end-to-end scheme. Compared with a traditional neural network, one of the advantages of CNN is the invariance of the spatial structure. This feature is much helpful for the segmentation task. In addition, in order to minimize the information loss for the one-to-one corresponding results, the main problem to be solved in deep learning based semantic segmentation task is how to scale the compressed feature maps back to the same size as the input image. E. Shelhamer et al. [13] propose to use transpose convolutional operations to perform up-scaling in the feature maps. Another work [14] utilizes the un-pooling operation to up-scale the feature maps by recorded positions of maximum value in max pooling. Formally, both of the works treat the segmentation CNN as “encoder-decoder” structures. In the encoding stage, the feature maps are extracted and downscaled by convolution and pooling operations. In the decoding stage, the feature maps are up-scaled and mapped to the specific dimensions. The various encoder-decoder architectures are widely used in the deep learning based image segmentation techniques such as instance and scene recognition or automatic driving tasks.

2.4 The deep learning techniques in the medical image processing area

With the fast developing of deep learning techniques in image processing domain, the usage of related methods in the medical image processing is also arouse attention by researchers and industrials. The classification of the disease of specific organs is the most directive usage of deep learning techniques in the medical image processing area. A. Esteva et al. [50] directly use the off-the-shelf network structure (Inception v3 [79]) on human skin images. Esteva et al. change the number of weights on the last output layer to the desired 2032 categories of skin cancers and pre-trained the model on nature image datasets. The results demonstrate good performance that outperforms the human doctors.

A similar approach also has been published based on the x-ray image. P. Rajpurkar et al. [15] utilize the off-the-shelf backbone network (DenseNet [80]) to classify up to 14 pulmonary diseases from raw x-ray images. In their works, the models are also pre-trained on nature images and they achieve remarkable results that surpass the professions in the related fields. Similar deep learning based classification works can be found in recent years such as M. Cicero et al. [81]. use Googe LeNet as the backbone network with the standard 224x224 input to achieve a classification model for abnormalities on frontal chest radiographs. G. M. Van et al. propose to use hard example mining to enhance the classification accuracy of the deep learning model for detecting hemorrhages in color fundus images [51]. The supports under these algorithms are a large number of labeled datasets and hardware accelerated training platforms.

Regarding the deep learning based applications for detection and segmentation in the medical image; some of the works make use of the 3D information by 3D CNN structure. For example by H. Chen et al/ in [82], the work proposes to use CNN structure with 3D convolutional kernels to detect bleed in brain MRI. The medical image has strong correlations in the sequential order; the 3D kernel takes advantages to this kind of dataset. However, the 3D convolution costs a large amount of computation and requires more memory space in the GPU device. J. M. Wolterink J M et al. use three independently trained CNN models with

shared structures from three perpendicular planes to detect the coronary artery calcium in 3D US image volume [83]. The network structure is designed as a full convolution structure without any pooling or fully connecting layers. The demerit of this kind of structure is a large amount of computation complexity. S. B. Lo et al. combine the template matching based method and deep learning model to detect the lung nodule from chest images [84]. The traditional template matching is adopted to extract all of the circle areas from the input image. Then, The CNN classification model is run on each of the candidates to judge if the area belongs to lung nodule. The method uses the CNN as a classification model, which cannot end-to-end learns from training dataset for detection task and causes high computation costs.

The existing deep learning based detection methods basically use the deep learning simply as the classification model. But as above-mentioned, the deep learning based object detector is more suitable for detecting multiple objects with different categories from given raw image input and it can provide end-to-end learning mechanism. The methods should be introduced to the medical image to improve the detection accuracy of the anatomical structure in US images. However, the positions of the detected bounding boxes are not so well aligned to the ground truth because of the errors of the regressed values and heavily affected by the imbalanced numbers of the positive and negative samples. Therefore, to make use of related works that applied in deep learning based object detector, this thesis proposes an optimized regression based CNN model to end-to-end learn the position of desired structures from raw US images.

In addition, some other works adopt the weakly-supervised method to infer the location of the interested object from learned weights of deep learning models. For instance, the above-introduced work [15] also provides the function to extract the discriminative map from the classification model. However, they do not provide the evaluation of the performance of the localization results. In addition, in the above-mentioned work [15], P. Rajpurkar et al. also provide the weakly-supervised localization results of the pulmonary diseases by cumulative activation mapping (CAM) [55]. However, the original CAM method extracts the area by finding the most discriminative area, which cannot represent the shape of the whole object. Therefore, the completeness of the region mining results still needs to be optimized. A similar

approach [54] tries to localize the different tissues in US image from the learned classification model of fetal standard planes. The feature localization method costs more computation and requires extra calculations on the back-propagations. One of the largest issues in the present weakly-supervised deep learning methods is lack of completeness of the segmented object area. The reason might be the classification model only focuses on the part of the object with the most discriminative features correspond to the feature level of output layers. For example, regarding the fetal head, the most discriminative area difference of the background object is the appearance of two ends of the fetal skull bones. However, the two ends of the skull cannot be seen as the entire area of the fetal head.

Regarding the pixel-wise classification of the anatomical structures or lesions in the medical image, in 2012, the authors of [52] proposed a deep belief network (DBN) [85] based segmentation framework to extract the area of left ventricle from US image, and they further improve the method in [86] for tracking the object in sequential US images. The method uses multiple cropped subregions and seeks for the edge of the target by pre-trained DBN network. They adopt DBN to un-supervised train stacked neural networks as a binary classifier to locate the most likely positions on the perpendiculars to the shape (contour) of anatomical structures. The works aim at segmenting the left ventricle, which is a single structure with a distinctive appearance in US image compared with the target of this thesis.

Other works use similar algorithms in the nature image as references. Such as [53], the authors choose to use the FCN [13] architecture to perform semantic segmentation of the left ventricle from US image. The feature maps are up-scaled to the original size by transpose convolution operations and iterative trained multiple segmentation networks in different output scales. The iterative training costs more time than the end-to-end trained model, and difficult to converge. The authors of another work [87] use symmetrically designed architecture to perform pixel-wise classification in US images. In their work, the authors use a similar optimization method proposed by K. –M. He et al in their network structure [9]. They add cross-layer connections to enhance the feature representation in higher levels. From the results of the papers, the segmented blobs still have space to be improved, especially for more

smoothed segmentation borders.

The above introduced related deep learning based approaches are summarized in Table 2.2.

2.5 Summary

As a summary, Section 2.1 and 2.2 first describe the problems of the traditional method in detection and segmentation anatomical structures from US images. The main problems existing in the traditional methods are: hard to fit the irregular shape by handcraft features, relies heavily on initial position, and cannot make use of global shape information from hard-to-defined features.

Throughout the related works in deep learning based works in other computer vision areas from Section 2.3, this thesis finds that the deep learning based technique might bring feasibilities to US image processing to achieve the targets of this research.

Section 2.4 makes a survey to the current existing deep learning based works that applied in related medical image processing area. Some of the work can be found that they still have space to further improvement. For example, in the segmentation of the anatomical structure, the existing methods cannot achieve enough accuracy on segmentation of fine-grained objects such as amniotic fluid and fetal body. In addition, the segmented blobs are lack of smoothness in the border of the object. On the other hand, some of the related works rely on pixel-wise annotations of the areas of the desired tissues, which cost much more manpower to make annotations. This thesis also proposes to adopt a weakly-supervised method to reduce the annotation workloads and improve the efficiency to the human doctors. According to the survey of the related works, even if there are some existing methods propose weakly-supervised approaches to solve the similar issues, their experiments demonstrate that the related methods are not good enough because of lack of completeness in the areas that do not have strong discriminative representations.

Chapter 3. Localization of Uterus

3.1 Introduction

As mentioned in Chapter 1, the location of the uterus plays an important role in the subsequent processes of automatic medical treatment systems. In addition, in this thesis, the following module makes use of the location of the uterus to suppress the data imbalance issue by the pre-generated region of interest. Therefore, as the first module, this chapter proposes a CNN based regression model to automatically locate the position of the pregnant uterus in raw US images.

The uterus is the soft tissue to carry the important anatomical structures such as fetus and amniotic fluid. The uterus has an inverted triangle shape and is located in the center of the pelvic cavity. Examples of fetal US images are visualized in Figure.3.1. The appearance of the uterus presents an irregular shape in US images, because of the gesture and posture of the fetus, and view angles etc. On the outer side of the uterus, the uterine wall can be observed. The uterine wall is composited by smoothed material but affected by noise and pseudo of the imaging devices, the appearance of the uterine wall is blurry in US images.

To localize the bounding box of anatomical structures in US images, conventional methods use manually designed feature descriptors and trained classifiers to detect the object in the images using a sliding window. For instance, N. B. Albayrak et al. [67] use “histograms of oriented gradients” (HOG) to extract the responses of the detection widow at each of the positions, and scores at each of the positions using a support vector machine (SVM) classifier so as to obtain the heat map of the anatomical regions in the US image. In Albayrak et al.’s explanations, they elaborate that their method requires prior knowledge about relationships between the relative positions of adjacent anatomical regions.

Figure 3.1 Examples of fetal US image.
This figure shows three different frames that are sampled from raw US video clips.

On the other hand, the deep learning (DL) based convolutional object detectors (such as Faster-RCNN [10] and SSD [11]) have recently shown large advantage over conventional methods for object detection by well-designed neural network structures. The DL based detectors map the input image into a target dimension space which corresponds to the vectors between sets of pre-defined positions to the ground truths. Inspired by their bounding box regression methods, this thesis makes use of the hieratically learned representations to achieve accurate uterine localization in fetal US images.

C. F. Baumgartner et al. [54] proposes a weakly supervised method which can localize the fine-grained tissues in US images such as fetal head and spine. Baumgartner et al.'s method localizes the region of interest in implicit obtained feature maps by learning inter-class distances. However, note that Baumgartner et al.'s method aims at learning the difference between image level annotations, which are not applicable to the task of this thesis, because most of the experimental data and annotations for uterus localization do not have such image level differences.

To verify the performance of deep learning based approach for uterus localization and improve the accuracy of the subsequent segmentation task, this thesis proposes a CNN based bounding box regression model to regress the offsets of a fixed number of multiple pre-defined positions (which is named as reference box).

Specifically, the proposed approach adopts an end-to-end learned regression CNN model. The input of this system is a still US image. The method feeds the raw US still image into a bounding box regression network to predict multiple candidate positions of the uterus. In

$$P = (X_{\mathrm{tl}}, Y_{\mathrm{tl}}, X_{\mathrm{br}}, Y_{\mathrm{br}})$$

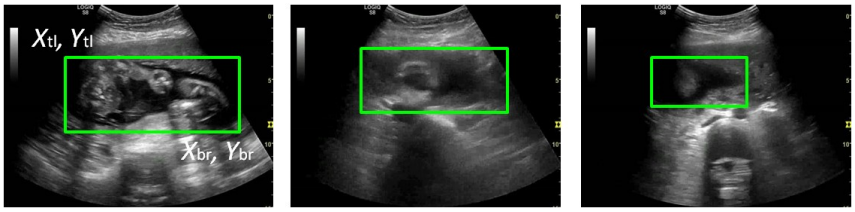

Figure 3.2 Definition of uterus bounding box.
Green rectangle indicates the ground truth bounding box of uterus.

order to do this, the bounding box regression network predicts the offsets between a set of pre-defined coordinates and the ground truth positions. The output vector of the network is a specifically designed structure which has a one-to-one corresponding relationship to each of the pre-defined positions. The element values of each position in the output vector indicate the offsets distance and probability that the uterus is at that position, respectively. During the testing, each of the predefined position is transferred by the predicted offset values and assigned as the predicted probability score. To obtain a final result from multiple candidate bounding boxes, the Non-Maximum Suppression [88] (NMS) is adopted as a post-processing approach to cluster the multiple positions into one position.

3.2 Methods

This section elaborates on how the proposed method localizes the position of the uterus in raw pregnant US images through the bounding box regression network. The overall framework, the regression structure, the backbone architecture, and the post-processing technique are detailed as follows.

3.2.1 Framework

The proposed method defines the position of the uterus as a tight bounding box, which can be represented by $P = (X_{\mathrm{tl}}, Y_{\mathrm{tl}}, X_{\mathrm{br}}, Y_{\mathrm{br}})$, where $X_{\mathrm{tl}}, Y_{\mathrm{tl}}, X_{\mathrm{br}}, Y_{\mathrm{br}}$ indicate the x (horizontal) and y (vertical) coordinates of the top left and bottom right corners, respectively (see Figure 3.2). The ground truth position P_{GT} of the bounding box is defined as the tight rectangle that starts

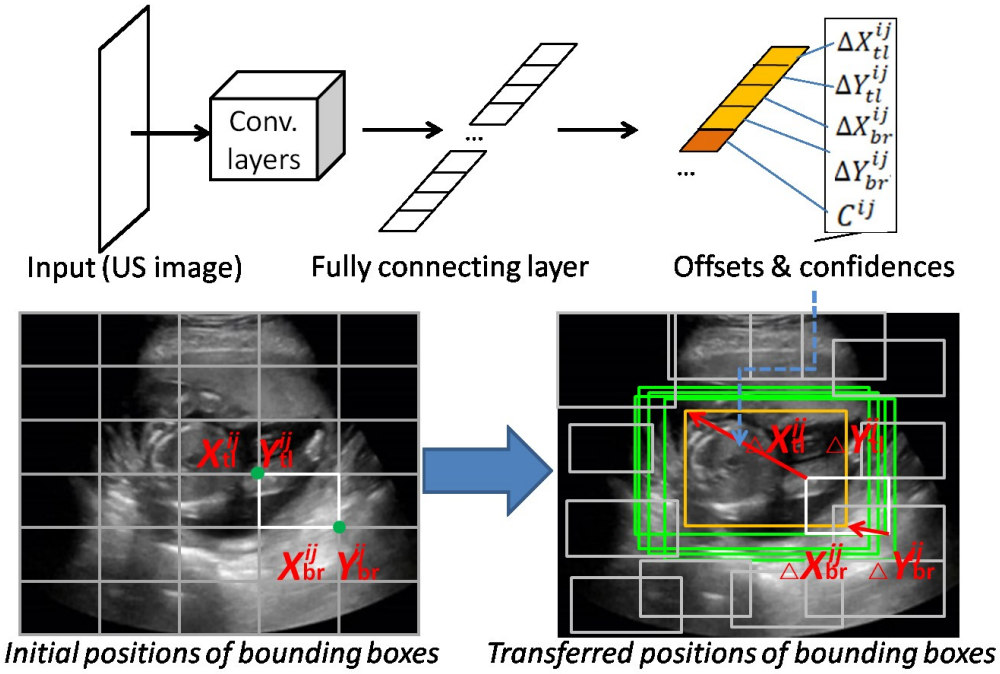

Figure 3.3 Framework of bounding box regression network.
Rectangles in green and orange: predicted bonding boxes.

from the leftmost pixel of the uterus, and ends at the rightmost pixel of the uterus. Similar to convolutional object detectors such as F-RCNN [10] and SSD [11], the proposed method uses multiple reference boxes with fixed initial positions and calculate the distance to the target positions. This thesis expects the deeply learned model to learn the offsets between the initial position and ground truth, and object confidence of each of the reference boxes as well. As a brief explanation of the proposed CNN regression model, the framework of the designed network and its output structure are illustrated in Figure 3.3.

3.2.2 Offset regression

To predict the position of the bounding box of the uterus by the convolution neural network model, this thesis designs a regression model which predicts the distance between sets of pre-defined locations and the ground truth locations. It defines multiple reference boxes through initial positions P_{Init}. For each sample image, we equally place n initial reference

boxes at every column and row. Given sample image $I(x)$, the offset is defined as $P_{\text{offset}}^{ij} = \left(\Delta X_{\text{tl}}^{ij}, \Delta Y_{\text{tl}}^{ij}, \Delta X_{\text{br}}^{ij}, \Delta Y_{\text{br}}^{ij}\right)$ for the reference box (i, j), where $(\Delta X_{\text{tl}}^{ij}, \Delta Y_{\text{tl}}^{ij})$ is the vector from the top left corners of the reference box to the ground truth, and $(\Delta X_{\text{br}}^{ij}, \Delta Y_{\text{br}}^{ij})$ is the vector from the bottom right corner of the reference box to the ground truth. Additionally, the localization model also needs to learn the confidence score C^{ij}, which indicates whether the reference box (i, j). The reference boxes is used as positive or negative samples by assigning different annotations (0 or 1) according to the overlapping area with the ground truth, as detailed below in the last paragraph of this section. Formally, the model learns the following mapping of $I(x) \sim [\begin{pmatrix} \Delta X_{\text{tl}}^{11}, \Delta Y_{\text{tl}}^{11}, \\ \Delta X_{\text{br}}^{11}, \Delta Y_{\text{br}}^{11}, C^{11} \end{pmatrix}, \ldots \begin{pmatrix} \Delta X_{\text{tl}}^{ij}, \Delta Y_{\text{tl}}^{ij}, \\ \Delta X_{\text{br}}^{ij}, \Delta Y_{\text{br}}^{ij}, C^{ij} \end{pmatrix}]$ for each of the reference box (i, j), where i and j range from 1 to n. All of the $(4 + 1) \times n \times n$ outputs are learned simultaneously by the weights of the network (Figure 3.3). This gives the method (normalized) predicted positions P_{Pred} for each of the reference boxes by

$$P_{\text{Pred}} = P_{\text{Init}} + F_{\text{d}}\big(I(x)\big), \tag{3.1}$$

where $F_{\text{d}}(U)$ indicates the trained bounding box regression network.

In this thesis' implementations, the offset value $\Delta\mathrm{X}$ and $\Delta\mathrm{Y}$ are normalized by the size of the original image, such as,

$$\Delta X = (X_{\text{Init}} - X_{\text{GT}})/W, \Delta Y = (Y_{\text{Init}} - Y_{\text{GT}})/H, \tag{3.2}$$

where $(X_{\text{Init}}, Y_{\text{Init}})$ and $(X_{\text{GT}}, Y_{\text{GT}})$ are the X and Y coordinates of the initial position and the ground truth, respectively; w and h are the width and height of the input US image, respectively. The normalization limits the output values so that the training losses converge.

The ground truth category C_{GT}^{ij} of each reference box is determined by the area V^{ij} in which the reference box and ground truth bounding box overlap, such as,

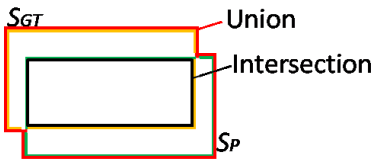

Figure 3.4 Illustration of IoU.
The area bounded by red indicates the union area, while the area bounded by black is the intersection area.

$$C_{\mathrm{GT}}^{ij} = \begin{cases} 0, \ if\ V^{ij} > \theta_{\mathrm{v}} \\ 1, \ if\ V^{ij} \leq \theta_{\mathrm{v}} \end{cases}, \tag{3.3}$$

where θ_{v} is a pre-defined threshold value (this thesis uses 0.3). The confidence labels of the corresponding reference boxes are set to 1 (positive); otherwise, the label is set to 0 (negative). The specific computation of the overlapping area V^{ij}, in which the ground truth bounding box and reference box overlap, are performed by Eq.(3.4),

$$V^{ij} = (S^{ij} \cap S_{\mathrm{GT}})/(S^{ij} \cup S_{\mathrm{GT}}) \ , \tag{3.4}$$

where S^{ij} indicates the area of the reference box (i, j), S_{GT} indicates the area of the uterus (ground truth) bounding box, $\cap$ and $\cup$ are intersection and union of the two areas, respectively (Figure 3.4). This thesis calculates the confidence value for each of the reference boxes at its initial position and concatenates them to the offset vectors. Note that in this implementation, all of the outputs (the sets of regression offsets and confidence scores) are put into one vector, which is together calculated as a linear combination by the fully connecting layer (Figure. 3.3). This vector based scheme can deal with robust global and context information, because all of the positions and confidence scores of the reference boxes are considered jointly.

Samples of labeled reference boxes are shown in the first column of Figure 3.5, where the rectangles in green and red are the positives and negatives, respectively.

3.2.3 Backbone network

This thesis uses the VGG16 (detailed in Table 3.1) structure as the base feature extractor

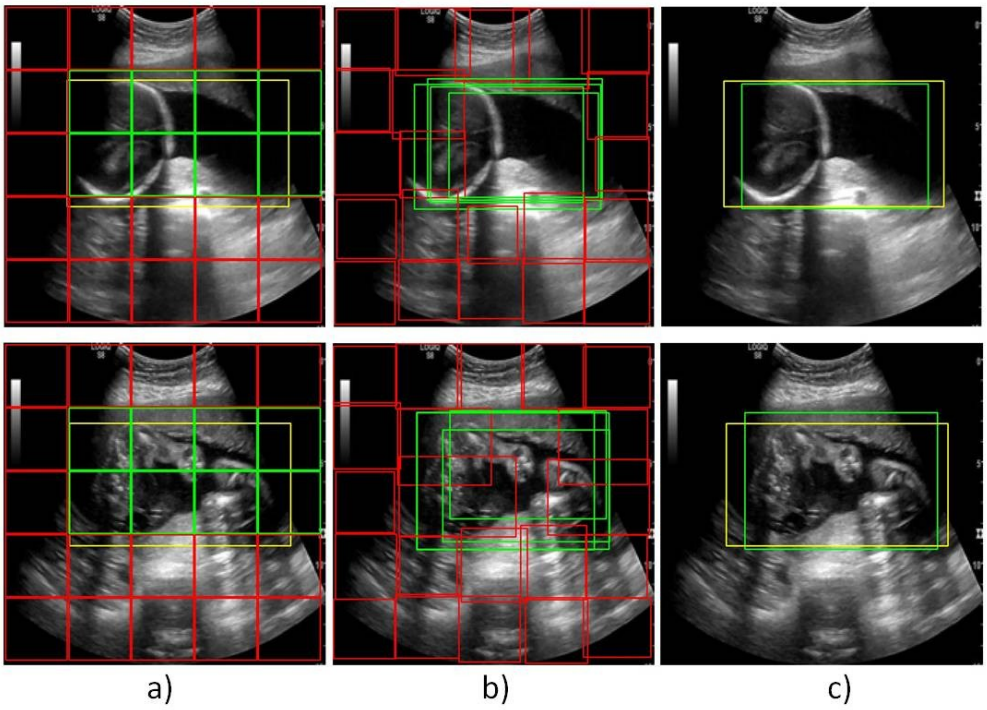

Figure 3.5 Initial position and transferred position of reference boxes. Column a) Initial reference boxes with category annotations (green: positive, red: negative) which are assigned by overlapping area with ground truth. Column b) The regressed positive (in green) and negative (in red) bounding boxes. Column c) The uterus localization results after NMS; the yellow rectangles indicate the ground truth bounding boxes. The images in upper and lower rows are two different frames that are sampled from raw US video clips.

(backbone network) for both bounding box regression and semantic segmentation networks. VGG16 has been successful in many computer vision and recognition tasks such as ImageNet [7] object classification challenge. Specifically, the feature extraction network is constructed by 14 convolutional layers. At the end of each of the linear combination operation this thesis uses Rectified Linear Units (ReLU) to perform an active function which maps the output feature to non-linear feature space. This thesis treats the backbone network as a feature extractor and does not measure effects on different backbone networks.

Regarding the structure of the output vector of the bounding box regression network, the model predicts the offsets and confidence scores C^{ij} for each of the $\mathrm{n} \times \mathrm{n}$ bounding boxes from per US image. As shown in Figure 3.3, this thesis organizes the output (the offsets and

Input: $P = \{p_1, p_2, \dots p_n\}, C = \{c_1, c_2, \dots c_n\}, \theta_{\mathrm{N}}$
P and *C* contain the initial predictions and corresponded confidence scores
P' and *C'* are the final predictions
$\theta_{\mathrm{N}}$is the preset threshold

P' = {}, *C'*= {}
while isempty(*P*)=*False* **do**
  i = argmax(*C*)
  P'.append(p_i), *C'*.*append*(c_i)
  P.remove(p_i), *C*.remove*(*c_i*)*
  for *j in P* **do**
    if iou(p_i, p_j) > θ_{N} **then**
      P.remove(p_j), *C*.remove*(*c_j*)*
    end
  end
end
return *P'*, *C'*

Figure 3.6 Pseudo code of NMS.

confidences) as a one-dimensional vector. Thus, the method converts the $n \times n \times 4$ offsets values and $n \times n \times 1$ confidence values to one long vector D_{pred} and use it as the output of the model. The Euclidean loss is adopted as the loss function of the detection network by,

$$L_{\mathrm{d}} = \| D_{\mathrm{pred}} - D_{\mathrm{GT}} \|^2 , \tag{3.5}$$

where D_{GT} is the ground truth vector. During the testing, this thesis feeds the raw US image into the bounding box regression network after resizing the input image to 224×224 pixels, which is same as the training phase. The network outputs the offsets and the uterine confidence of each of the reference boxes in float numbers.

3.2.4 Post-processing

As introduced earlier, the method first transfers all of the initial coordinates to the target positions P_{Pred} by corresponding offsets, as indicated by Eq. (3.1). The initial prediction contains many overlapped bounding boxes which are assigned to the same object in the image (as shown in Figure 3.5 (b)). Therefore, the approach needs to further eliminate redundant

predictions. This thesis clusters the multiple overlapping boxes by Non-Maximum Suppression (NMS).

The NMS seeks for the position with the maximum confidence value in a given region which might contain the same object by eliminating all of the bounding boxes whose overlapping areas are larger than a threshold. The pseudo code of the common NMS method used in this thesis can be found in Figure 3.6. Consequently, the merged bounding box is obtained as the final result of the uterus detection. Among the obtained multiple candidate bounding boxes, this thesis keeps the bounding box which has the largest confidence as per NMS.

3.3 Numerical Results and Discussions

This thesis aims at providing solutions for real-world problems; therefore, all of the methods proposed in this thesis are evaluated using clinical dataset. First, detailed experiment environments and training parameters are explained. Then, to verify the effectiveness and seek for the optimal parameters for further modules, numerical results are obtained through experiments. This thesis discusses the performance in several different settings and gives intuitive conclusions through visualized examples.

3.3.1 Experimental environment

To conduct experiments, this thesis receives approval from the Ethics Review Committee on Research with Human Subjects, Waseda University (2014-165). The examinations use GE Voluson E8 and C1-5 linear array transducer with frequencies in the range of 4.0 Hz. The axial and lateral resolution is 2mm and 3mm, respectively. The field of views is 66 degrees and the depth setting is 15cm. The average fetal age of the subjects is approximately 22 weeks. This thesis acquires four video clips from anonymous patients. The original resolution of each frame is 640×480 pixels. The raw data are sampled from each video clips every two frames. This thesis uses 2-fold cross validation by first randomly separating the different patients' data into the training and testing sets. The contour annotations are made by doctors who have years of US examination experiences.

Input: $I = \{I1, I2, \ldots In\}$,
$F()$ is the feature extraction model
T_{Clean} is the cleaned dataset
θ_p is the preset threshold

$T_{Clean} = \{\}$
$Ic = I1$
T_{Clean}.append (Ic)
for *i in n* **do**
  if *Edist(F(Ic),F(Ii))>thrs* **then**
    $Ic = Ii$
    T_{Clean}.append (Ic)
  end
end
return T_{Clean}

Figure 3.7 Pseudo code of data cleaning method.

3.3.2 Data cleaning

Because of the manual operation of the antenatal examination, the ultrasonic probe moves at non-uniform velocity during the scanning, and sometimes stops for a period of time. If frames are directly taken at equal sampling interval from the video data, the experiments get a large number of repetitive or nearly duplicated frames. The duplicate or nearly duplicate images do not contribute to the training, and cause bias to the dataset. The bias could affect the distribution of the training data and the data diversity, which results in bad generalization to the regression models. Therefore, this thesis adopts a simple and effective method to filter out highly similar samples between adjacent frames in the dataset. Actually, the method one by one feeds each of the data samples through a pre-trained CNN model (i.e. Alexnet [7] network structure and trained on ImageNet dataset). Then the feature vector outputted from the fully connecting layer is adopted as the decomposed representation of the input image. The extracted feature vector of the current frame is compared with the next frame in the entire image sequence. If the Euclidean distance of the current frame and previous frame is smaller than threshold, the previous frame is discarded. Till the method finds an image which has large enough distance with current frame the model update the current frame to this one. The

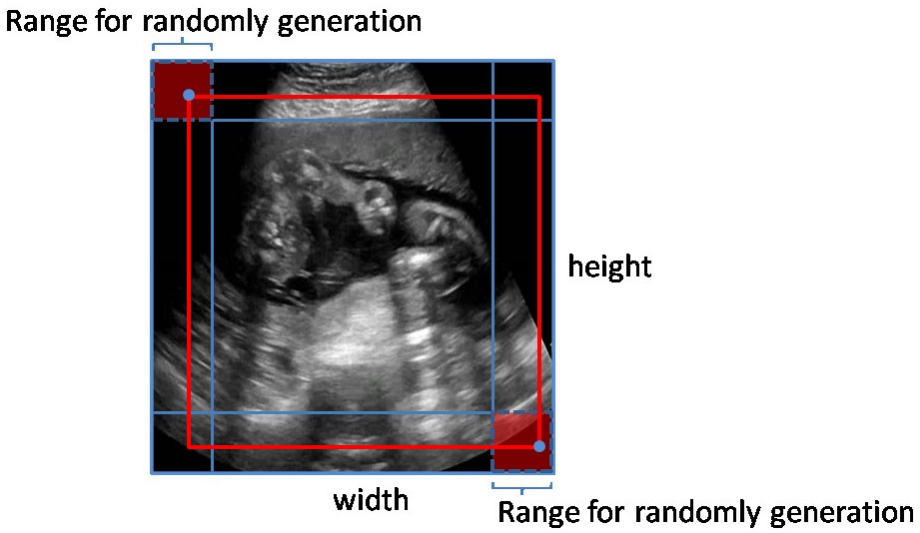

Figure 3.8 Data augmentation method.
The area in the red rectangle is used as the augmented image for training.

pseudo code of the data cleaning method is shown in Figure 3.7. In other words, image pairs which have a very short distance are treated as the similar image samples. This thesis discards all of the redundancies from the original data set both for training and validation. The cleaned dataset has about 400 samples in total for training and validation.

3.3.3 Data augmentation and hyper parameters

In order to enhance the diversity of training data, the method augments the training data by adding random disturbances to the predicted uterus bounding boxes. In particular, multiple sub regions are cropped from the original image with random transfer factor on the four points of the bounding box. Detailed parameters and illustration for augmentation method are shown in Figure 3.8. The data augmentation increases the training samples to more than 4000 training samples in each subset.

The experiments run on a single NVIDIA GTX 1080. The deep learning platform is modified from Caffe [89]. For each iteration, the method feeds multiple images into one batch and minimizes the error of the entire batch. The error between the ground truth and output is summed and averaged over all of the samples. This thesis stops model training after 20000

Table 3.1 VGG16 backbone network.

Conv. Layers of VGG16
Conv1_1: 64x3x3, Stride: 1, Pad: 1
Conv1_2: 64x3x3, Stride: 1, Pad: 1
Pool1: 3x3, Stride: 2
Conv2_1: 128x3x3, Stride: 1, Pad: 1
Conv2_2: 128x3x3, Stride: 1, Pad: 1
Pool2: 3x3, Stride: 2
Conv3_1: 256x3x3, Stride: 1, Pad: 1
Conv3_2: 256x3x3, Stride: 1, Pad: 1
Conv3_3: 256x3x3, Stride: 1, Pad: 1
Pool3: 3x3, Stride: 2
Conv4_1: 512x3x3, Stride: 1, Pad: 1
Conv4_2: 512x3x3, Stride: 1, Pad: 1
Conv4_3: 512x3x3, Stride: 1, Pad: 1
Pool4: 3x3, Stride: 2
Conv5_1: 512x3x3, Stride: 1, Pad: 1
Conv5_2: 512x3x3, Stride: 1, Pad: 1
Conv5_3: 512x3x3, Stride: 1, Pad: 1
Pool5: 3x3, Stride: 2

iterations. The inference time on GPU is about 30ms for the detection and 60ms for the segmentation networks per image, respectively.

3.3.4 Domain transferred learning

Studies proved that the low level feature representations are highly similar across many domains. Compared with randomly initialized weights, the fine-tuning on the pre-trained model converges at a faster speed and achieves better performance. The weights of the convolution layers of the segmentation model are initialized to a pre-trained model [48]. Note that the only difference between training from randomly initialized weights and training from pre-trained weights is the starting (initial) values of weights of each learn-able layer. In addition, only the operations which have exactly same number of weights are adopted from pre-trained weights. Therefore, during fine-tuning of the models, the newly added layers are initialized by Gaussian distributions and have five times as large learning rate as the other

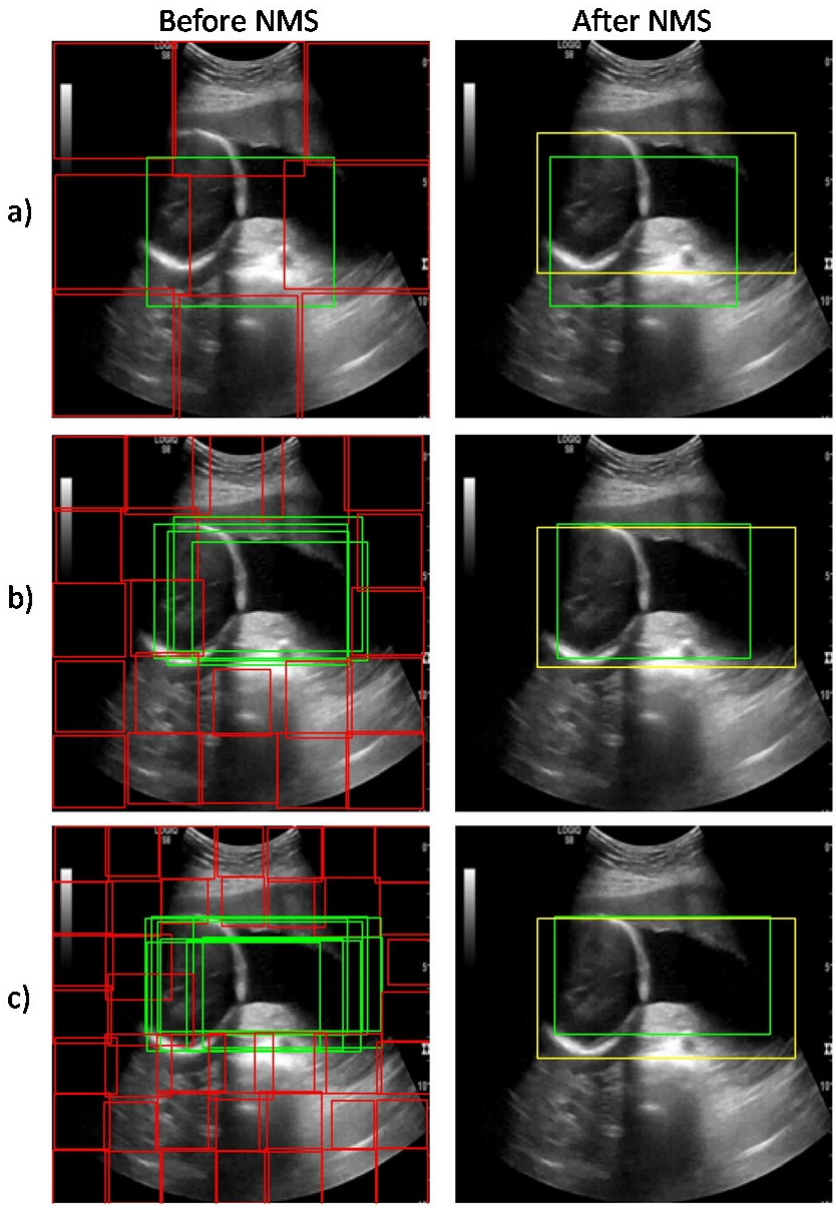

Figure 3.9 Comparison between a) 3x3, b) 5x5 and c) 7x7 reference boxes (green: positive, red: negative, yellow: ground truth), left column: bounding boxes before NMS, right column: bounding boxes after NMS.

fine-tuned layers.

3.3.5 Evaluation criteria

The experiments use intersection of union (IOU) to evaluate the uterus detection networks.

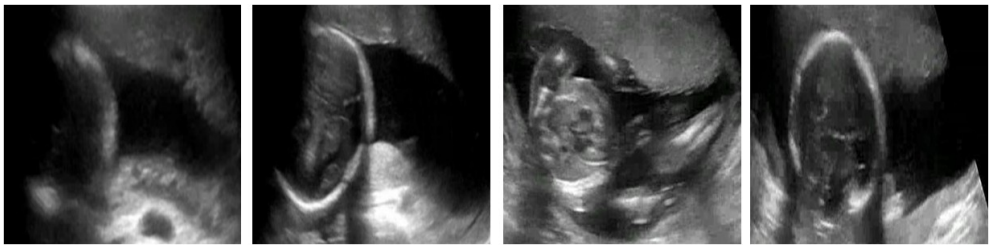

Figure 3.10 Normalized uterine images by detected bounding box.
The figures from left to right show four different frames that are sampled from raw US video clips.

Table 3.2 Evaluation results of uterus localization (IOU) (%)

	URN_3x3	URN_5x5	URN_7x7	FRCNN	SSD
Sub.1	55.7	63.1	**64.6**	61.0	60.7
Sub.2	55.0	**61.7**	59.7	60.2	59.1
Avg.	55.3	**62.4**	62.1	60.6	59.9

Table 3.3 Evaluation results of uterus localization (AP) (%)

	URN_5x5	FRCNN	SSD
AP.5	**99.6**	99.5	99.4
AP.6	**88.6**	85.4	83.3
AP.7	**68.3**	53.7	50.2

Specifically, IOU is the proportion of the intersection of true positive pixels and union area of the predictions and ground truths. Besides IOU, the experiments also adopt Average Precision (AP) [97] to evaluate the detection performance. To demonstrate the different alignment accuracy between ground truth bounding box and prediction, this thesis uses three overlapping thresholds: 0.5 (AP.5), 0.6 (AP.6) and 0.7 (AP.7). The bounding boxes which have overlapping ratio larger than the thresholds are considered as True Positives (TP). It means that the larger threshold is stricter than smaller one.

3.3.6 Results and discussions

This section compares different settings of the bounding box regression network. The quantitative results of IOU and AP are shown in Table 3.2 and Table 3.3, respectively. The proposed uterus regression networks in this thesis are named as URN.

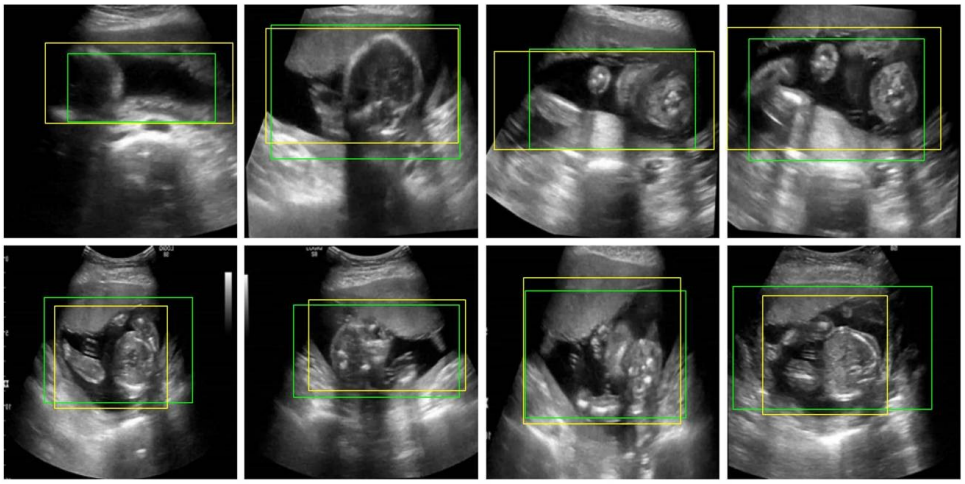

Figure 3.11 Visualized results of uterus detection.
Green: manual annotation; yellow: prediction. This figure shows eight different samples.

This thesis evaluates the uterus localization results on different setting and compares them with Faster-RCNN [10] (FRCNN) and SSD [11]. For FRCNN this thesis uses multi-scale training scheme to randomly resize the input image in three different scales. During testing this thesis uses the original size as input. Regarding SSD, this thesis adopts 300x300 pixels as both training and testing inputs. For both of the above-mentioned methods, this thesis adopts VGG16 as the backbone network.

The IOU shows that URN_5x5 achieves the best score but the gaps between different results are not obvious. The AP demonstrates that URN_5x5 model achieves the highest accuracies, especially in case of more strict criteria (AP.7). It proves that the proposed method has better alignment accuracy. The major difference is the last fully connecting operation. It introduces global feature to the offsets regression, which brings richer context information. The global regression scheme is more robust in case of uterus localization in US images.

Regarding the different settings of the reference boxes, as shown in Figure 3.9, the comparisons of multiple different settings by using 3x3, 5x5 and 7x7 boxes as the initial positions can be intuitively observed. This thesis observes that in most of the situations the predictions have smaller size than the ground truth in width. This is caused by the unclear area at the two sides of the sector. Although the predicted uterus area is smaller than the

uterus (caused by unclear shadows), it does not affect the subsequent scheme too much. Since the method extends every bounding box with a fixed factor, this issue mostly happens on the left and right sides of the US image, which contain relative large black area.

Through comparing different settings on the output structures, the results show that the output positions which are generated with 3x3 reference boxes has relatively smaller overlapping areas with the ground truth bounding box, which causes lower IOU as a result of comparing with 5x5 and 7x7. The results prove that the model with more dense reference boxes achieve better performance on localization accuracy. This is because in the preliminary experiment, the uterus area has relatively large size, compared with the original image; the effectiveness of using more densely regressed positions cannot be well reflected with this dataset. On the other hand, inaccurate localizations are concentrated on the left and right side on x axis. It is hard to fit to the target bounding box if the two sides of the object are heavily cast by shadow in the US image; the edge of the uterus is unclear and out of the range of the view angles. In contrast, the top and bottom position of the predicted bounding box on y axis is relatively accurate, and this result is enough for this thesis to further optimize the subsequent modules, because most of the background areas are positioned in the upper and lower area of the pregnant US images.

The advantages of using cropped areas can be explained in terms of the following two aspects: 1. constrain the location of the uterus to the area inside the uterus bounding box; 2. alleviate the unbalanced data distribution. It can be said that the predicted uterus area smaller than the uterus because of unclear features at the left and right areas (some of the visualized results are shown in Figure 3.11). This does not heavily affect the following segmentation scheme, because the method extends every bounding box with a fixed factor. As a pre-processing stage of anatomical structure segmentation, the uterine localization results relieve the severe data imbalance problem caused by background pixels. The performance of the localization model can be viewed from the quantitative results using IOU. The IOU reflects the alignment accuracy of prediction and ground truth. The best result achieves higher than 62% on averaged IOU with ground truth. Regarding the detection rate, the proposed

model achieves acceptable accuracy (higher than 88% at AP.6) for the uterus detection task, and outperforms other compared existing deep learning based approaches which are applied to nature image domain.

3.4 Conclusion

This chapter has proposed a deep learning based uterus localization frameworks. The method adopts specifically designed regression output structure to regress candidate positions of the uterus. In particular, to obtain the abundant position information of the target object in US images, multiple densely positioned reference boxes are assigned according to the size and length-width ratio of the original image. The target of the model is to learn the offsets between each of the reference box and target position and predicts the confidence of the positive sample. In order to achieve this, the output of the network is designed as a vector which has the same length as the coordinates and confidence of all of the reference boxes. During the training, the model is optimized by minimizing the loss between predictions and manually annotated ground truth.

Through experiments this thesis verifies the methods and concludes the results as follows: the detection rate of the uterus using the best the model achieves about 88% for AP@IOU>0.6 using clinical pregnant US dataset, and higher than 62% for IOU averaged with the ground truth. The proposed US uterus localization scheme outperforms other compared deep learning based methods, which are applied for nature image domains. The results demonstrate the proposed method can achieve relatively high detection performance on predictions with almost no miss detections. In this thesis, this module works as the beginning of the subsequent processes. In the next chapter, to improve the performance of the semantic segmentation of anatomical structures, the approach makes use of the results estimated by this uterus regression CNN.

Chapter 4. Semantic Segmentation of Anatomical Structure

4.1 Introduction

As mentioned in Chapter 1, the anatomical structures inside the uterus such as the fetal body and amniotic fluid reflect several important physiological indexes of the pregnant patients. Segmenting anatomical structures in US images is very important for achieving many automatic systems such as fetal biometry measurement, 3D fetus reconstruction, and computer-aided amniotic fluid test. In the last chapter, a uterus localization CNN is proposed; the position of the uterus provides a region of interest for fine-grained structure segmentation. Therefore, this thesis further proposes a method to distinguish the areas of anatomical structures in the given US images.

The segmentation is challenging, as shown in Figure 4.1, because of irregular shapes of the uterine walls deformed by the fetal bodies that could change their postures, noisy reflected waves (black shadow-like areas in Figure 4.1) that are generated by tissues and bones, and blurry edges. On the other hand, the structure density based segmentation relies on low-level features to distinguish the liquid and tissues from US examinations. It is difficult to distinguish the different tissues, and cannot obtain smooth segmentation on the border of the object.

Energy function based boundary searching method is one of the common ways to learn deformable shape information in US images: for example, P. R. Thangaraj et al. [34] adopt a watershed-based method to use seed regions for identifying the areas of renal calculi in US images. However, these methods either heavily rely on strong and clear image patterns or

Figure 4.1 Examples of pregnant US image.
This figure shows three different frames that are sampled from raw US video clips.

require extra precise initialization; therefore, it is difficult for these methods to be applied to the pregnant uterine US image segmentation.

Other than a low level feature, extracting the edge of the object is commonly used to segment the object in a closed region. N. Martins et al. [90] adopt the energy function and statistic model-based approaches. Martins et al. adopt an active contour model to fit the morphable model to the target shape (contour) by minimizing the energy functions which are defined on the perpendiculars to the key points on the contours. B. Georgescu et al. [91] propose a data-driven approach to detect and segment the edge of the left ventricle (LV) in US images. Georgescu et al. first detect the desired object by hand-craft features and boosted cascade trained classifier. Then Georgescu et al. use selected features to obtain the shape inference on the aligned LV images. Georgescu et al.'s method is not directly applicable to the setting, because it is only capable of binary classification and it relies on regular shape information.

Apart from these model-driven or pre-defined rules-based approaches, deep learning based machine learning mechanism achieved great success in nature image processing. As a data-driven method, convolution network structures could approximate any objective function by tons of neurons and effectively choose the most discriminative features by local receptive fields filtering on input image signals. Recent research works have shown that CNN (Convolutional Neural Network) could handle segmentation on complex scenarios such as road images and indoor object images [13][14][92], by treating the problem as a pixel-wise classification task. They have shown that deep learning could get better results than traditional approaches from automatically learned hieratical representation and relative large training

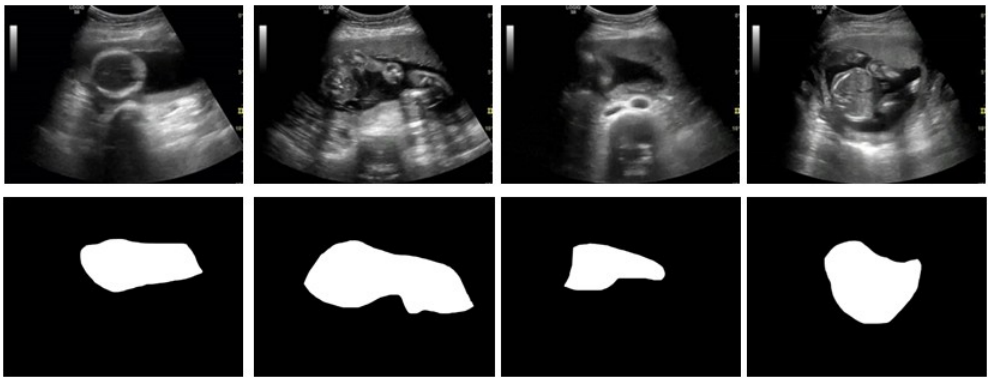

Figure 4.2 Examples of pixel-wise annotation of uterus.
1st row: four original pregnant US image samples; 2nd row: corresponded ground truth annotations (in white) of uterus.

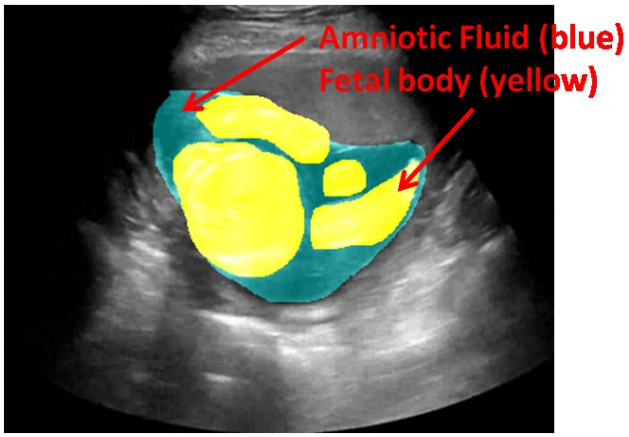

Figure 4.3 The illustration of the desired anatomical structures.
Amniotic fluid (in blue), fetal body (in yellow) and other areas are considered as background.

data is better than one or more manually designed features for semantic segmentation of objects from nature images.

Regarding the deep learning based US image segmentation, G. Carneiro et al. [52] adopt a deep belief network (DBN) to train stacked neural networks as a binary classifier to localize the most likely positions on the perpendiculars of the contour of anatomical structures. The feature is extracted on each of the positions of the anchor points to obtain the responses, which makes redundant computations in overlapping areas. Their annotation model, however, is not stable when applied to a unique shape which is not included in the training. H. Chen et

al. [53] used off-the-shelf fully convolution network (FCN) structure [13] to segment the LV in designated US slices. They iteratively segment the desired object in the subregion of the US image. The results show that the segmentation in the detailed areas of complex scenes still needs to be improved.

This thesis explores the effectiveness of applying the deep learning technology to segmenting pregnant uteruses in US images. More specifically, this thesis aims at automatically determining whether each pixel belongs to the desired anatomical structures or not in input pregnant US images. The segmentation methods proposed in this thesis are separated into two parts: the first part is the binary segmentation of the uterus in US images, and the other part is the multi-category segmentation of amniotic fluid and fetal body in US images. For the binary segmentation of the uterus, the definition of the "uterus" area is the pixels inside the uterine walls, where the uterus area could probably include the amniotic fluid and fetal body. The rest of the pixels are defined as "non-uterus", as shown in Figure 4.2. Regarding the semantic segmentation of amniotic fluid and fetal body, the fine-grained structures are pixel-wise labeled inside the uterus, as shown in Figure 4.3.

To solve the above two problems, in following sections, this thesis first introduces a fully convolution network structure which outputs the binary confidence that each pixel corresponds to the uterine or background area in the input image. In the training phase, the weights of the network are updated using manually labeled training data. During the test phase, this thesis directly feeds the entire image to the trained model. Then, this thesis extends the method into multi-categories segmentation of amniotic fluid and fetal body. The output structure is modified, and the target of the learning is changed to multinomial loss function. The experimental results show that the segmentation in the detailed areas of complex scenes still needs to be improved. In section 4.4, optimization methods are introduced to relieve the "inaccuracy" issue and "unsmoothed" segmentations on the object borders. By conducting experiments that use pregnant uterine US images, the effectiveness of the proposed methods is explored.

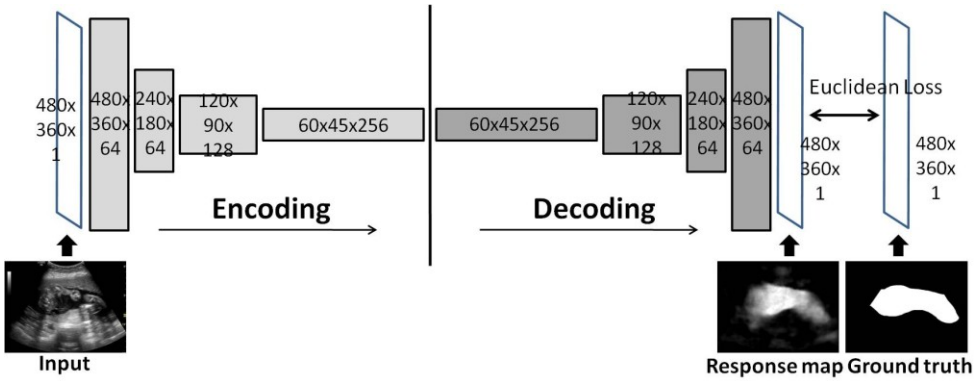

Figure 4.4 Encoding-decoding network structure for binary segmentation of uterus. Light gray: Conv-RELU max pooling in the encoding part; Dark gray: up-sample conv in the decoding part.

4.2 Preliminary research: Semantic segmentation of Uterus

This thesis first designs preliminary research to segment the uterus area in the US image. More specifically, this thesis adopts an off-the-shelf deep learning segmentation framework to verify the effectiveness in US images.

The overall framework of the preliminary research is shown in Figure 4.4. Detailed explanations are as follows.

4.2.1 Network structure for binary segmentation

The main difference from common CNN with full connected layer structure is that the output layer of the CNN this thesis uses is a dense map, which is a set of multiple labels. For this, this thesis uses an encoding-decoding CNN structure and pixel-wise loss which is designed for object segmentation, as shown in Figure 4.4.

The network consists of two parts: encoding and decoding parts. Specifically, in the encoding part, all of the kernel sizes for the convolution layer is fixed. The method pads each response map by zeros on the border of the matrix (kernel) in order to keep the size fixed after being filtered by the convolution kernel. Every convolution layer is followed by a rectified linear unit (Rely [94]) activation function for the models. Then, the size of the response maps is

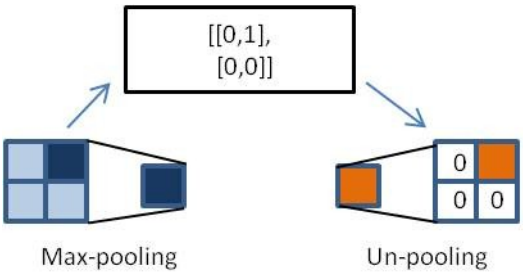

Figure 4.5 Down-scaling and corresponded up-sampling operation.

determined as the one obtained by decreasing the size of the pooling layers. This could yield scale invariance and represent information from surrounding (larger) areas. For the pooling strategy the model uses max pooling which shrinks the input map by only keeping the maximum value of the specific pixel neighbors. Both of the size and stride this thesis uses for each pooling layer are 2. Note that the method needs to record the max indices, in order to transfer back the response map by the up-sample layers in the decoding part of the network. The method repeats this kind of layer blobs (conv-RELU-maxpooling) four times totally in the encoding network.

Concerning the decoding part, to decode the encoded response maps without losing dense (the same resolution as the input) information, the model needs to up-scale the layer to the same resolution as the input image. This could be achieved by several optional up-scale schemes, for example by using bi-linear transform [13], or learn-able de-convolution kernels [92]. This method records maximum position to up-sample the feature maps to larger scale in the decoding stage (see Figure 4.5). In particular, the max pooling indices are recorded in the encoding layer as mentioned earlier, so that for each up-scaled image, the model could restore the pixel values by finding the corresponding values and put them in the previously recorded position. The other pixels are all filled by zeros. Then, each up-sample layer is followed by the convolution layers with activations as usual as described before. The method uses exactly same numbers of layers in the reversed order as the encoding layers.

For each input image this thesis resizes it to a fixed size of 480*360 pixels, and converts it to a single gray level channel. Since the method uses a fully convolution network structure, there

are no fully connecting layers, which allows us to handle any arbitrary input resolution in the testing phase, and the output is always an end-to-end segmentation map. The specific resolution of each layer's output is shown in Figure 4.4. Different kernel numbers are set for each layer; for the first two convolution layers there are 64 shared kernels for each and for the following two layers the kernels are increased by factor of square of 2. For the decoding part the model uses corresponding number of kernels as shown in Figure 4.4. The last layer has only one layer output, which indicates the confidence of uterus or non-uterus at each pixel.

To deal with the segmentation task for the uterus area, the binary segmentation CNN directly compares the Euclidean distance with the ground truth in training. In particular, the training procedure calculates the loss of the network by summing each sample's pixel-wise Euclidean distance between the response map and ground truth label map. The loss function of the segmentation network is shown as:

$$L_s = \frac{1}{2}\sum_{x\in \mathrm{I}} \left\| v_{\mathrm{g}}(x) - v_{\mathrm{p}}(x) \right\|_2^2 \tag{4.1}$$

where x indicates a pixel in image I, $v_g(x)$ is the ground truth label and $v_p(x)$ indicates the outputted pixel value at pixel x. During the back propagations, for the decoding part of the network, the derivatives are down sampled by recorded max indices. This process is performed in the opposite direction of the decoding process.

To summarize, the method uses an encoding-decoding CNN network structure to perform an end-to-end, pixel-level supervised learning mechanism. Compared with only using down-scaled feature map output and then up-scaling by interpolations, the usage of up-sample layers followed by convolution layers brings dense predictions for each of the positions that correspond to every pixel on the original input.

4.2.2 Result thresholding

The value of each pixel of the response map indicates a confidence value of whether it belongs to "uterus" or not. In order to obtain a reasonable threshold for binary mask for the

uterus class despite imbalanced numbers of pixels in the two classes (uterus or non-uterus), before setting this threshold to the testing set, the model first runs on training set. By testing several different thresholds within a specific range, a best threshold could be determined. This thesis assumes the testing set has a similar distribution with the training set. The method treats this threshold value as the most suitable threshold during the testing.

4.2.3 Experiments

Dataset The experiment uses a GE Voluson E8 and C1-5 linear array transducer with frequencies in the range of 4.0 Hz. The average fetal week of the subjects is approximately 22 weeks. Concerning the parameters of the data capture device, axial and lateral resolution is 2mm and 3mm, respectively, and the field of views is 66 degrees. The depth setting is relatively high (15cm), because the entire transverse section of patient's uterus should be observed in the US slices. During the clinical examination, the doctors scan from side to side on the entire abdomen several times for each of the patients. The gestational weeks of the fetal video used for this experiment are around 19 and 23 weeks. As a preliminary experiment, 226 frames for the training set (week 19) and 188 for the test set (week 23) are sampled from the videos. Both of the training and testing images are resized to a size of 480*480 pixels in order to directly input to the CNN.

Each pixel's label is manually given: specifically, if a person specifies the (enclosed) area of the uterus in each US image using a graphic tool, then, the "uterus" label is given to all the pixels inside the specified uterus area, while the "non-uterus" (background) label is given to the other pixels. Note that for each US image, it is ensured that in almost all of the samples the fetal body could be observed inside the uterus.

Metrics The method uses following three different metrics for quantitative evaluations of the results: 1. Global accuracy (Accu_G): the properly classified pixel counts divided by the total pixel counts in the dataset. 2. Mean accuracy over all categories (mAccu_C): averaging the pixel-wise classification accuracy of all of the classes (in case of this preliminary experiment, two classes). This index could reflect the imbalanced factor of each class. 3. Mean

Table 4.1 Overall segmentation accuracy. (Preliminary Exp.) (%)

	Accu_G	mAccu_C	mIoU
EuclidieanL_k3	90.45	84.99	73.14
EuclidieanL_k5	94.28	90.05	82.12
MultinomialL	93.00	90.46	79.57
EuclidieanL_k7	**95.20**	**90.73**	**84.42**

Table 4.2 Class separated segmentation accuracy. (Preliminary Exp.) (%)

	Accu_BG	Accu_U	IoU_BG	IoU_U
EuclidieanL_k3	93.17	76.80	89.05	57.23
EuclidieanL_k5	96.39	83.71	93.35	70.89
MultinomialL	94.27	**86.66**	91.82	67.32
EuclidieanL_k7	**97.42**	84.05	**94.41**	**74.43**

intersections over union (mIoU): the IoU of each class i is calculated by the following equation:

$$\mathrm{IOU_i} = (S_{pi} \cap S_{gi})/(S_{pi} \cup S_{gi}) \ , \tag{4.2}$$

where S_p indicates the area of predicted pixels, while S_g indicates the area of corresponding ground truth pixels. Equation (4.2) indicates that the IoU reflects both of the false positives and the false negatives; therefore, it can be said that this criterion is more strict than the other two criteria.

Training details Concerning the hyper parameters of the model, the method uses a fixed learning rate at $1*10^{-6}$. The learning rate is quite small, because the loss is calculated by summing all the pixels of the image, which could generate a large value for the loss. The loss tends to decrease if it reaches approximately 16,000. This thesis stops the training at around 500, which corresponds to 20,000 iterations or more. Under this network structure and these settings, the training program costs about 6,600 Mb GPU memories with batch size set to 4. It takes for the training about 5 hours for 20,000 iterations on a NVIDIA GTX1080.

4.2.4 Numerical Results and Discussions

First, whether the assigned label is correct or not is checked at each pixel in all of the test data.

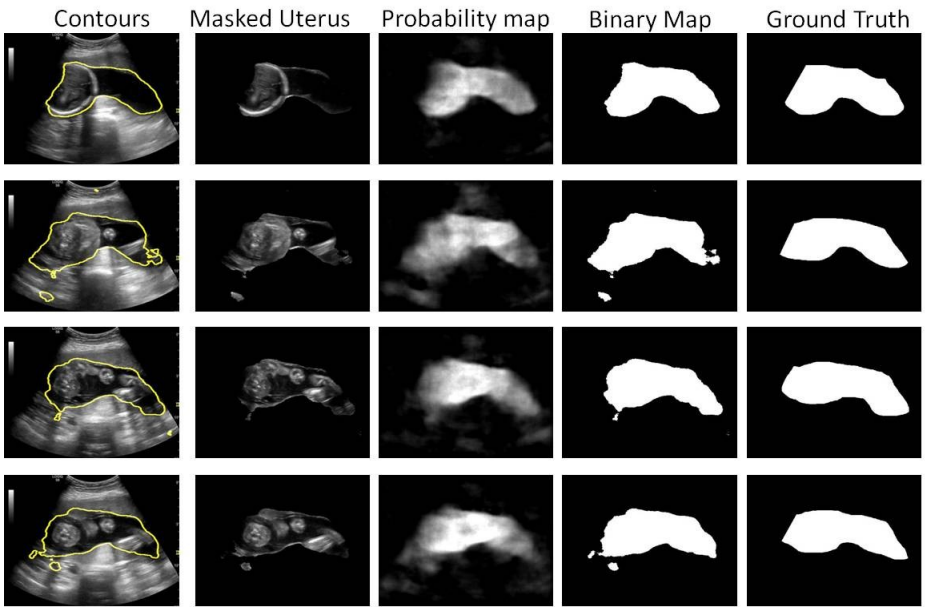

Figure 4.6 Down-scaling and corresponded up-sampling operation.
Each row shows one different US image sample.

Table 4.1 lists the three criteria for the Euclidean loss (EuclidieanL), where Euclidean_k3, Euclidean_k5 and Euclidean_k7 are the results of using different kernel size 3x3, 5x5 and 7x7 pixels, respectively. As can be seen in Table 4.1, among the three different kernel sizes, the largest kernel size (7x7) gives the best segmentation accuracy.

Second, the segmentation accuracy is evaluated in each of the two classes. To evaluate this accuracy, Accu_BG, Accu_U, IoU_BG and IoU_U are used, which indicate the accuracy and IOU of the background (_BG) and uterine area (_U). From Table 4.2, it can be said that the background (_BG) area gives higher accuracy than the uterus (_U). Since the number of pixels in the non-uterus is larger than the uterus, their pixel value variation is larger; thus, many pixel values tend to be treated as the background. Another reason is that the features of the fetal body in the uterus tend to be quite similar to those of the background area.

Some examples of the result are visualized in Figure 4.6. In Figure 4.6, the four columns correspond to four different US images, and four rows show the original US image, ground truth for the uterus area (indicated by white), the obtained border between the uterus and non-uterus (yellow line) and the obtained uterus area (white), respectively. The segmentation

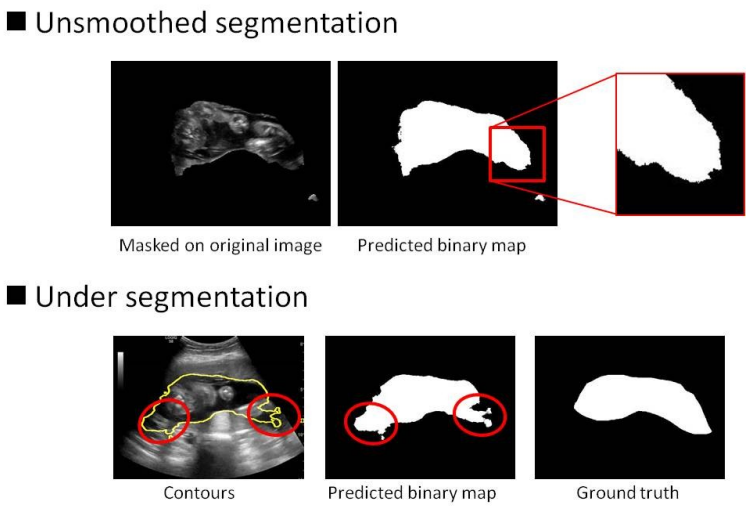

Figure 4.7 Major issues in uterus segmentation results.
Area in red rectangle shows the unsmoothed contour. Areas in red circles indicate the under segmentation results.

is hardly affected by fetal bodies, which tend to have very similar pixel value variations in non-uterus pixels. This indicates that the CNN based method is robust to the easy-to-be-confused local patterns.

4.2.5 Issues

On the other hand, some problematic cases can be seen. The research shows the major issues in Figure 4.7. The problems can be classified into the following cases. First, the wrong segmentations occur in areas contaminated by noisy reflections, specifically, the misclassified pixels easily appear at the blurry edge (unsmoothed segmentation). In such an area, enough amounts of training data was not obtained, and a large size of the convolution kernel cannot cover this problem. Second, the under segmentations are obtained. Errors often happen in case that part of the border (typically, the left and/or right border) of the amniotic fluid area in the uterus overlaps with the outer (out of the field of view of US probe) area, where the gray-levels of amniotic fluid and outer area are dark and similar. In addition, in order to achieve better results for future applications, this method still needs to be further improved, especially on the object borders.

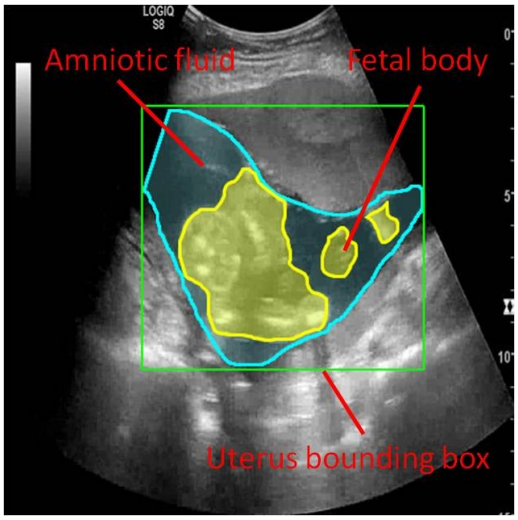

Figure 4.8 Example of the fetal US image and its annotation.
The green rectangle indicates the bounding box of uterus.

4.3 Optimized multi-categories semantic segmentation

In the above-mentioned sections a method for binary segmentation of the uterus in US images is described. However, in real world solutions, systems require for more fine-grained information for areas of multiple structures. In addition, from the results of the preliminary experiment, there are defects in some aspect such as the in-accurate segmentation and un-smoothed borders. Therefore, this section further extends the research to multi-object segmentation by multi-category segmentation CNN and proposed several optimization methods to solve the issues.

To achieve this goal, this thesis proposes a two-tier approach of deep learning based techniques to (I) locate the bounding box of the uterus and (II) segment the pixels in the uterine region into fetal body, amniotic fluid, and background. An example of fetal US image and its fine-grained annotation of amniotic fluid and fetal body are visualized in Figure 4.8.

First, the method makes use of the localized uterus area to improve the semantic segmentation results. The segmentation can be constrained in the area of the uterus, and the cropped image relieves the data imbalance issues that caused by background area.

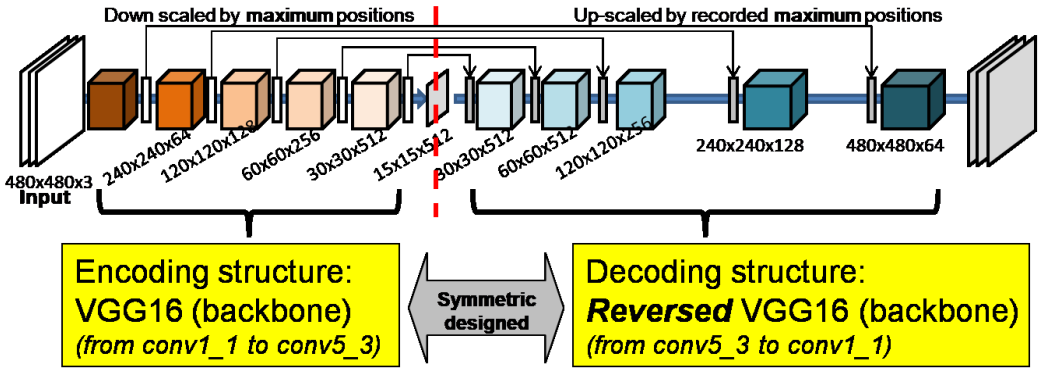

Figure 4.9 Symmetric designed encoding-decoding framework.

Other than segmentation in the region of interest, this thesis proposes an optimized CNN based architecture to segment the fetal body and amniotic fluid in an end-to-end, fully supervised learning pipeline. To further improve the predictive performance in complex scenes and smooth out the segmentation results, this thesis modifies the structure with additional layers and multi-scale supervisions.

In particular, regarding the base framework of the segmentation CNN, for each input image the method directly feed the data blob to a down-scaled convolution structure, so as to obtain multiple down-scaled feature maps. Then, corresponding un-pooling layers followed by learn-able convolution kernels are adopted to decode the feature maps to the same size as the input image. To deal with the multiple category output, the output structure is changed to multiple output channels which indicate the confidence maps for different anatomical structures. First, the basic technique elements of the module are described as follows.

4.3.1 Encoding decoding framework

As shown in Figure 4.9, the method still follows the encoding-decoding framework which is same as the preliminary experiments. The input image is first mapped to a set of down-scaled feature maps by convolution and pooling operations. This work uses max-pooling, which keeps only the maximum value of each pooling window. In the first half stage of the network (encoding stage), multiple max-pooling operations are adopted. Overall, the feature maps of the last layer of the encoding stage are down scaled to $(1/2)^{\mathrm{M}}$ of the raw input by M

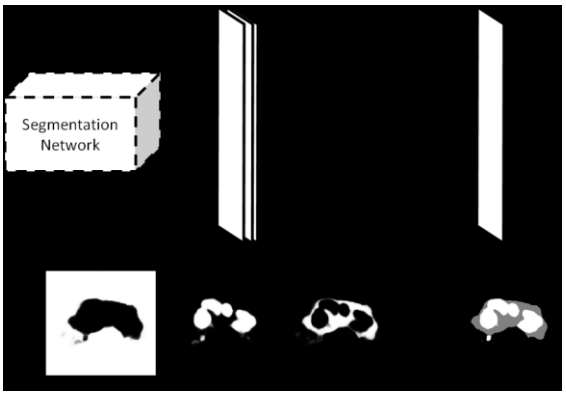

Figure 4.10 Pixel-wise softmax for multi-category segmentation.

max-pooling layers with stride two.

In the encoding stage, the input image is mapped to a set of down-scaled feature maps by convolution and pooling operations. Normally, in case of other image processing tasks such as object classification, the dimension of the feature maps is further reshaped to one dimension feature vector, and then sent to the output layer. However, concerning the image segmentation task, if the model directly connects the down-scaled feature map to the output, the information owned by the neighbor pixels in each of the pooling windows is lost. Therefore, to achieve a pixel-to-pixel classification without losing dense information, the method needs to up-scale the feature map back to the same size of the input image by learn-able weights. There are several ways to resize a matrix such as bi-linear interpolation and nearest neighbor. Here, this thesis adopts un-pooling [93] to restore feature maps to a larger size, which is same as the corresponding feature maps in the encoding stage. As shown in Figure 4.5, each pixel is assigned to the recorded maximum position in each of the corresponding windows, while the other positions are filled with zeros. For the backward propagation, the derivatives are passed to the former layer by only keeping the largest one in the window, which is same as the feed forward operation in the max-pooling layer. For each un-pooling layer, the model concatenates them with convolution layers which have learnable convolution kernels. In the overall architecture, it has numbers of un-pooling layers as the max pooling layers. In the last layer of the decoding stage, the feature maps are finally

up-scaled to the original size. To up-scale the feature maps to the exactly same size as the input image, the parameters of the un-pooling layers one-to-one correspond to the pooling layers in the encoding stage.

4.3.2 Backbone network

Concerning the network architecture, the method uses the first 10 convolution layers of VGG16 [48] as the base network structure. The network uses sets of convolution kernels with relatively small (3x3) convolution kernels. This model uses four max-pooling layers with stride of two pixels. This thesis fixes the input image size to 480x480 pixels, where the smallest size of the feature map is 15x15 (the last convolution layer in the encoding structure). Note that this module treats the backbone network as feature extractor and do not measure the effects on different backbone networks. The specific resolution of each layer's output is demonstrated in Figure 4.9.

In order to predict the confidence maps of objects of multiple categories, the method extends the output structure to multiple channels and changes the loss function to multinomial logistic loss.

During testing, the method calculates the softmax at each pixel of the last output map in channel dimensions (as shown in Figure 4.10). During training, for each input image I with Q pixels, the corresponding loss function L_{ms} for the proposed multi-category segmentation network is:

$$L_{\mathrm{ms}} = -\frac{1}{Q}\sum_{x=1}^{Q} \log p_x \ , \tag{4.3}$$

where p_x is the predicted confidence of pixel x which corresponds to its ground truth category. The error between the ground truth and output is summed and averaged over all of the pixels of each batch. The author uses stochastic gradient descent (SGD) method to update weights in the back propagation training.

Note that this architecture does not have any fully connected layers. Therefore, structurally,

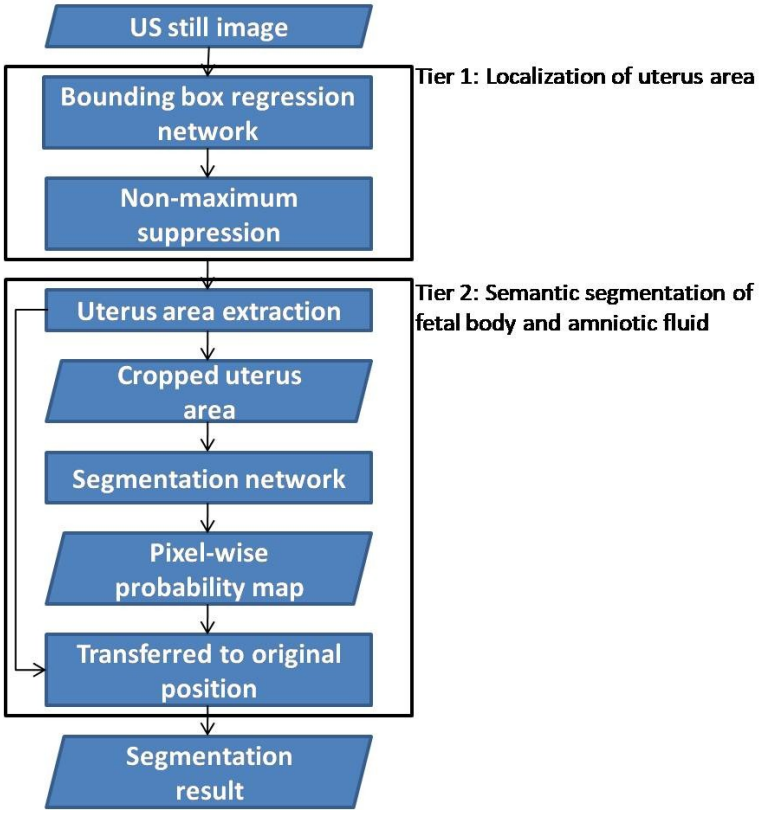

Figure 4.11 The flow map of proposed 2-tier approach for multi-category object segmentation.

the input does not need to be fixed for each of the convolution layer. Even so, during the training and testing phases the model uses fixed size as the input, in order to make the local receptive field keep consistent.

4.3.3 Optimized Semantic Segmentation Framework

Through the preliminary experiments, several issues have been revealed. First, the segmented blobs lack of global representation, which cause the in-accurate segmentation on some objects with irregular shapes. Second, the experiments reveal that the model is hard to converge in training and final segmentation results tend to miss-classify many of the pixels into background areas because of the imbalance distribution between negative (background) and positive (anatomical structures) pixels. Third, the segmented blobs lack smoothness in the border of the object, which generates rough edges in the reconstructed models.

In order to avoid the issues mentioned above, this thesis proposes an optimization framework for the segmentation of anatomical structures in US images. First, to relieve the data imbalance issues caused by large area of background pixels, the optimized method makes use of the result of uterus localization by CNN to shrink the input area. What is more, to improve the segmentation accuracy, the method proposes to use addition 1x1 sized kernels to enhance the global representation of the segmentation model. In addition, to solve the problem of un-smoothed segmentations, this thesis further proposes a multiple intermediate supervision method to achieve smoothed segmentation on the border of the objects. Details of each building block of this research are introduced as follows. The overall working flow of 2-tier framework is illustrated in Figure 4.11, and the detailed the optimizations structures are shown in Figure 4.12.

4.3.3.1 2-tier segmentation

The input of the system is a still US image. The method first feeds the raw US still image into a bounding box regression network (see Chapter 3) to localize the position of the uterus. Then, it extracts the uterus area by the obtained bounding box and then resizes the sub region of the image to a fixed size. An optimized segmentation CNN is used to classify each of the pixels to pre-defined categories in the input region. Finally, the pixel-wise probability map is transferred back to the origin position by the localized uterus area.

As shown in Figure 4.11, this section describes the semantic segmentation model for anatomical structure segmentation. To prevent that the obtained bounding box is smaller than the actual uterus area, the method extends the bounding box area with a fixed factor of 1.2 to leave some residuals for the areas near the uterine border.

4.3.3.2 Inner layers

The higher layers of the network structure represent more high level responses. To this end, this thesis adds layers called “inner layers” between the encoding and decoding stages to enhance high level representations (as shown in the middle part of Figure 4.12). The inner layers are stacked 1x1 sized convolution kernels which are connected to the last layer of down

sampling network. The inner layer works as smallest size local receptive fields. It does not do integral on local regions, but it shifts the dimensionality in filter space. With the increase in the network parameters, the feature space of the model becomes larger. To this end, the model directly uses more layers with learnable weights to enhance the hieratical representation. For each of the inner layers, the model uses much more kernels than any other layers in order to map the features to relatively large dimensions. Here, 1x1 sized kernels do not significantly increase the computation cost compared with layers with larger kernels.

The local appearances of human tissues are very similar to each other. The feature of general objects such as human and cars, local appearance has stronger discriminative power in gradient, color, texture, etc. However, it is hard to parse objects from partial features in US images. The false alarm tends to happen at similar areas with adjoining edges if the model lacks global representations. To verify the effectiveness this thesis compares several settings with different inner layers; details of the experimental results are presented in Section 3.

4.3.3.3 Intermediate supervision

In the decoding stage of segmentation network, it can be said that the un-pooling operation causes un-smoothed segmentation in the border areas. By comparing with interpolation up-scaling, the un-pooling operation's output is relatively sparse, because most of the positions are filled with zeroes. It causes the final segmentation results to be very non-smooth in some areas, especially in the boundaries.

As shown in the output part of Figure 4.12, the method proposes to improve the smoothness by using multiple output branches with multi-scaled supervision signals. The images with different resolutions yield different detailed information in the borders. The higher resolved segmentation maps contain more information in the border area and have smoother edges. Thus, this thesis expects to enhance the continuity of each group of convolution layers by different resolved ground truth segmentation maps. Specifically, inspired by GoogLeNet [8], the model not only uses the error from the last output layer, but also from inserted additional output branches among the decoding structure. The additional outputs are used as auxiliary

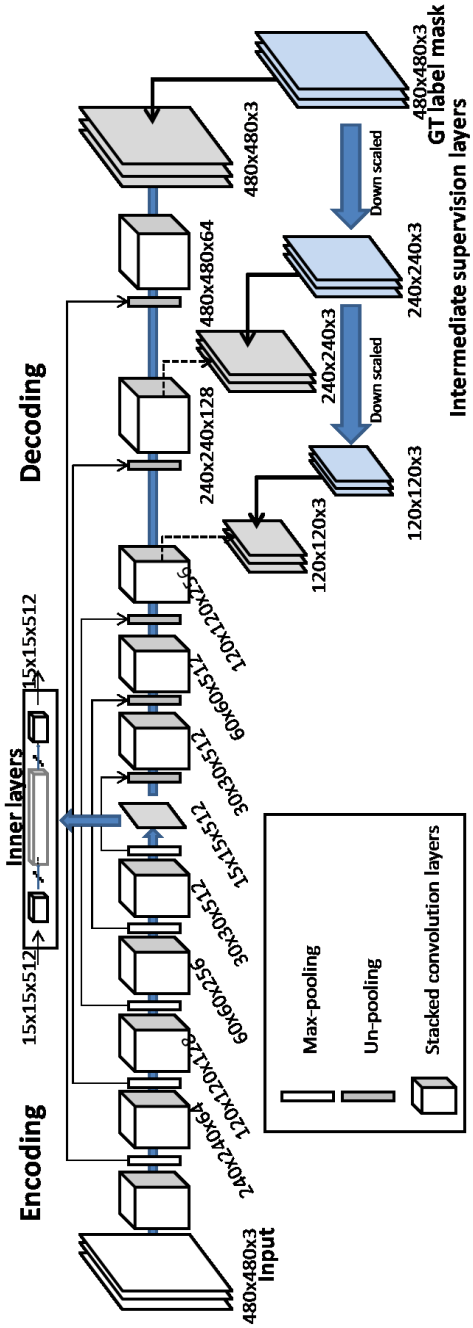

Figure 4.12 Optimized semantic segmentation framework for semantic segmentation of multi-category anatomical structures.

branches for calculating the errors between the predictions and down-scaled ground truth label maps, they are concatenated to the output of convolution layers in the up-sampling stage.

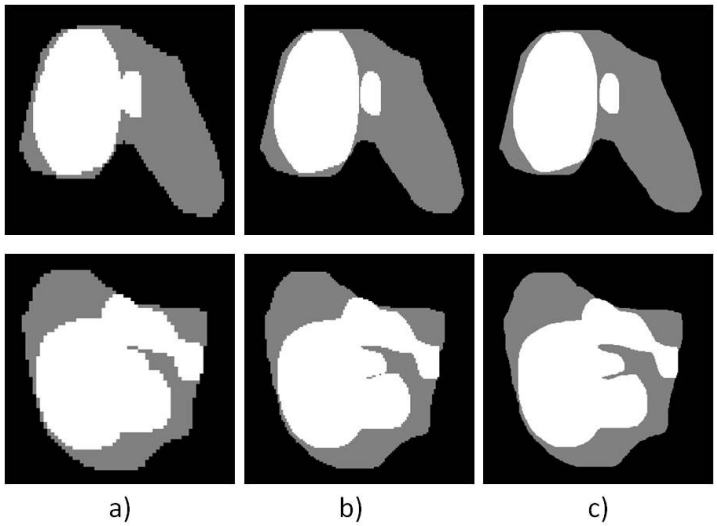

Figure 4.13 Ground truth label maps with different image scales.
a) 120x120, b) 240x240 and c) 480x480 sized output branches. (The images are resized to the same size for visualization). Each row indicates one different ground truth of US image sample.

To map the intermediate output to the same dimension number as the ground truth, for each of the intermediate branch the method uses convolution layers with a fixed number of convolution kernels (the number of kernels equals to the number of channels of the output).

Here, this thesis uses nearest neighbor interpolation to down-scale the ground truth label map from the original size to the same size as the corresponding output branch. Examples of down-scaled label maps are visualized in Figure 4.13. The corresponded resolution of the label maps are 480×480, 240×240 and 120×120 pixels. Through intermediate supervision, the convolution layers after un-pooling tend to learn extra information from small to large segmentation maps. The larger sized segmentation maps contain more information in the border area. In Figure 4.13, the difference between the different sized segmentation maps are mainly in the border of the blobs. The larger sized segmentation maps have smoother edges. Thus, the proposed method expects to enhance the continuity of each group of convolution layers after un-pooling operations by learning the differences between multi-sized ground truths.

During training, each of the intermediate supervision layer targets at minimizing the cross-entropy loss of all of the positions in the response map. This thesis treats all of the errors equally by summing the derivatives of each of the branches with same weights during the back propagation training. The method does not weight the error of each branch layer by the size of the output map, because the loss is normalized by the number of pixels of the output.

During testing, the auxiliary branches are discarded, and only the output of the last layer is used as the final segmentation results.

4.3.4 Experiments

The evaluations are conducted using clinical US image dataset, which has approval from the Ethics Review Committee on Research with Human Subjects, Waseda University (2014-165).

Data acquirement The data source used in this verification is same as the preliminary experiments, which are introduced in Section 4.2.3. The method re-samples the US frames in the raw US images to obtain more samples for training and testing. As a result, there are in total over 900 images for the training and 400 images for the testing (three subjects for training and the last one for testing).

To obtain the ground truth annotations, a radiologist manually labels each of the US images in the cleaned dataset. In order to automatically estimate the amniotic fluid volume and fetal body size by segmented blobs, there are in total three categories (fetal body, amniotic fluid and background) need to be annotated. The contour annotations are made by doctors who have years of US examination experiences. The pixel-wise labels of each category are assigned by judging to which closed region each pixel belongs, as shown in Fig.4.13. Note that for each image, it is ensured that one could observe the uterus in all of the samples.

Data cleaning In addition, as introduced in Chapter 3, this thesis eliminates completely duplicate or nearly duplicate frames in the original sampled data. These duplicate or nearly duplicate images do not contribute to the training and increase the manpower for label

annotations. The feature is first extracted by pre-trained CNN model, and then the experiment discards the similar samples from the original data set. The finally obtained clean dataset has 413 for the training set and 188 for the testing set. The segmentation network is equivalent to a pixel-wise classification task. Although the dataset is relatively small, each pixel could be seen as a training sample.

Metrics Besides intersections over union (IOU), this thesis further demonstrates the class specified evaluation results by ROC (Receiver Operating Characteristic) curve. The true positive rate (TPR) and false positive rate (FPR) over classes are calculated by:

$$\mathrm{TPR} = \mathrm{N_{TP}}/\mathrm{N_{Pos}}, \mathrm{FPR} = \mathrm{N_{FP}}/\mathrm{N_{Neg}}, \tag{4.4}$$

where $\mathrm{N_{TP}}$, $\mathrm{N_{FP}}$, $\mathrm{N_{Pos}}$, $\mathrm{N_{Neg}}$ indicate the number of true positive, false positive, positive and negative pixels of each category, respectively. This thesis also adopts pixel-wise classification accuracy (Accu) to evaluate the segmentation results without the effect on different categories. Note that the segmentation results are calculated based on the original map. The outside area is judged as background.

Training details The experiments run on a single NVIDIA GTX 1080. The deep learning platform is modified from Caffe [89]. The error between the ground truth and output is summed and averaged over all of the samples. The training is stopped after 20000 iterations. The method uses a stepped learning rate starting from $1\mathrm{x}10^{-2}$, and decreases the learning rate by a factor of 0.1 at each of 4,000 iterations. Each of 50 iterations takes about 51 seconds in the training phase. The inference time on GPU is about 60ms for the segmentation networks per image.

Data augmentation The method uses same augmentation method as introduced in Chapter 3. It makes the training samples increase to more than 4000 training samples in each subset. For comparison, the result w/o data augmentation ("_w/o_aug") can be found in Table 4.3.

Domain transferred learning The above-mentioned experiments have proved that the low level feature representations are highly similar across many domains. Therefore, this method initializes the weights of the convolution layers of the segmentation model from pre-trained

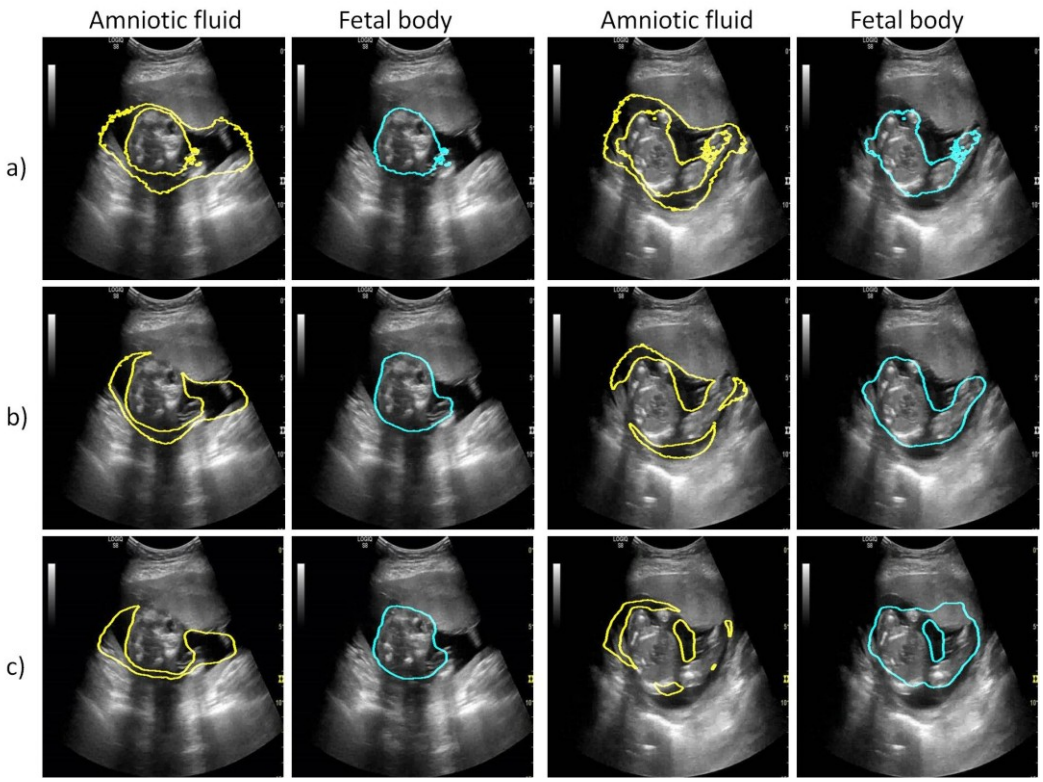

Figure 4.14 The visualized segmentation results.
a) SegNet (w/o intermediate supervision layers), b) EDN_IS (w/ intermediate supervision layers) and c) PSPVGG16.

model [48]. During fine-tuning of the models, the newly added layers have five times as large learning rate as the fine-tuned layers. For comparison, the result w/o using pre-trained model ("_w/o_pre") can be found in the last row of Table 4.3.

4.3.5 Numerical Results and Discussions

This thesis names the optimized scheme "EDN_". The base scheme without any modification is same as Segnet; thereby, this thesis names the baseline work of EDN "Segnet". To demonstrate the performance of intermediate supervision layers, in Fig.4.14, this thesis shows some of the segmentation results by highlighting the edge of the segmented blobs. The smoothed boundaries which are predicted by model with intermediate supervision layers show significant improvements in the given image samples. The intermediate supervision layers also improve the overall segmentation accuracy (EDN_IS of Table 4.3). The visualized

Table 4.3 Accuracy over all of the pixels (Accu). (%)

	Segnet_ nodet	Segnet	EDN_IS	EDN_1IL _IS	EDN_2IL _IS	EDN_3IL _IS
w/ cropped uterus area		✓	✓	✓	✓	✓
w/ intermediate supervision (_IS)			✓	✓	✓	✓
w/ Inner1 (_1IL)				✓	✓	✓
w/ Inner2 (_2IL)					✓	✓
w/ Inner3 (_3IL)						✓
Subset 1	91.99	92.23	93.43	93.97	93.81	**94.08**
Subset 2	90.41	91.05	92.22	92.39	92.49	**93.71**
Avg.	91.20	91.64	92.83	93.18	93.15	**93.90**

Table 4.4 Structures of inner layers used by different models

Layer name	EDN	_1IL	_2IL	_3IL
Layer1	None	512	4096	4096
Layer2	None	None	512	4096
Layer3	None	None	None	512

Table 4.5 Class specified results (IOU) and pixel-wise accuracy over all of the pixels (Accu). (%)

	Fetal body	Amniotic fluid	Bkg.	mIOU	Accu
FCN8S	49.32	36.06	92.37	59.25	89.86
SegNet	55.00	44.83	93.35	64.39	91.64
PSPVGG16	65.13	**55.13**	94.84	71.70	93.03
V3+_VGG16	69.17	52.64	95.19	72.33	93.51
EDN_IS	64.10	51.71	94.90	70.24	88.01
EDN _w/o_aug	44.29	30.91	92.03	55.74	89.90
EDN _w/o_pre	61.24	50.81	93.10	68.38	92.83
EDN_3IL_IS	**69.36**	54.79	**95.44**	**73.19**	**93.90**

results (Figure 4.14) show significantly improvements especially on the border of the segmented areas.

Regarding the 2-tier segmentation framework, it can be said that the predicted uterus area is

smaller than uterus, it is caused by unclear features at the left and right areas (as shown in Figure 3.11). It does not heavily affect the segmentation scheme, because this thesis extends every bounding box with a fixed factor. The quantitative results of semantic segmentation w/ and w/o cropped uterus area can be found in the column 1 ("SegNet_nodet") and 2 (SegNet) of Table 4.3. It proves that the localization scheme improves the accuracy of the following segmentation work.

Then, to verify the effectiveness of the inner layers, this thesis uses several settings with different number of the inner layers. The detail of the structures of the inner layers is listed in Table 4.3. The suffix "_nIL" indicates the model with n (n=1, 2, 3) inner layers. The visualized results (Figure 4.15) show the additional inner layers bring higher accuracy on the segmentation results. Table 4.5 shows the class specified evaluation results by IOU for amniotic fluid and fetal body. It shows that the additional learn-able weights in the middle of the encoding decoding network lead to better performance among other models. The detailed pixel-wise classification results can be viewed from the category specified ROC curves, which are drawn in Figure 4.16.

Besides this, the "_w/o_pre" (no pre-training) is trained without domain transfer learning and has same network structure to "EDN_3IL_IS". In addition, the "_w/o_aug" (no augmentation) is trained without data augmentation and has same network structure to "EDN_3IL_IS".

Concerning comparison experiments, the research adopts FCN and Segnet as baseline methods. Besides the baseline methods, this thesis also evaluates Deeplabv3+ [95] and PSPNet [96] with same training and testing sets. Note that all of the segmentation models are trained based on the cropped uterus area for fair comparison. The FCN model adopts transpose convolution to learn the up-sample operations. The FCN8s model is iteratively trained by FCN32s and FCN16s. Regarding the Deeplabv3+, the major feature of the method is that it adds multiple atrous convolution kernels (ASPP) with batch normalizations to the last feature maps to enlarge the field of views. This research adopts the ASPP scheme with batch normalizations and an additional encoder-decoder layer. Regarding the PSPNet, this

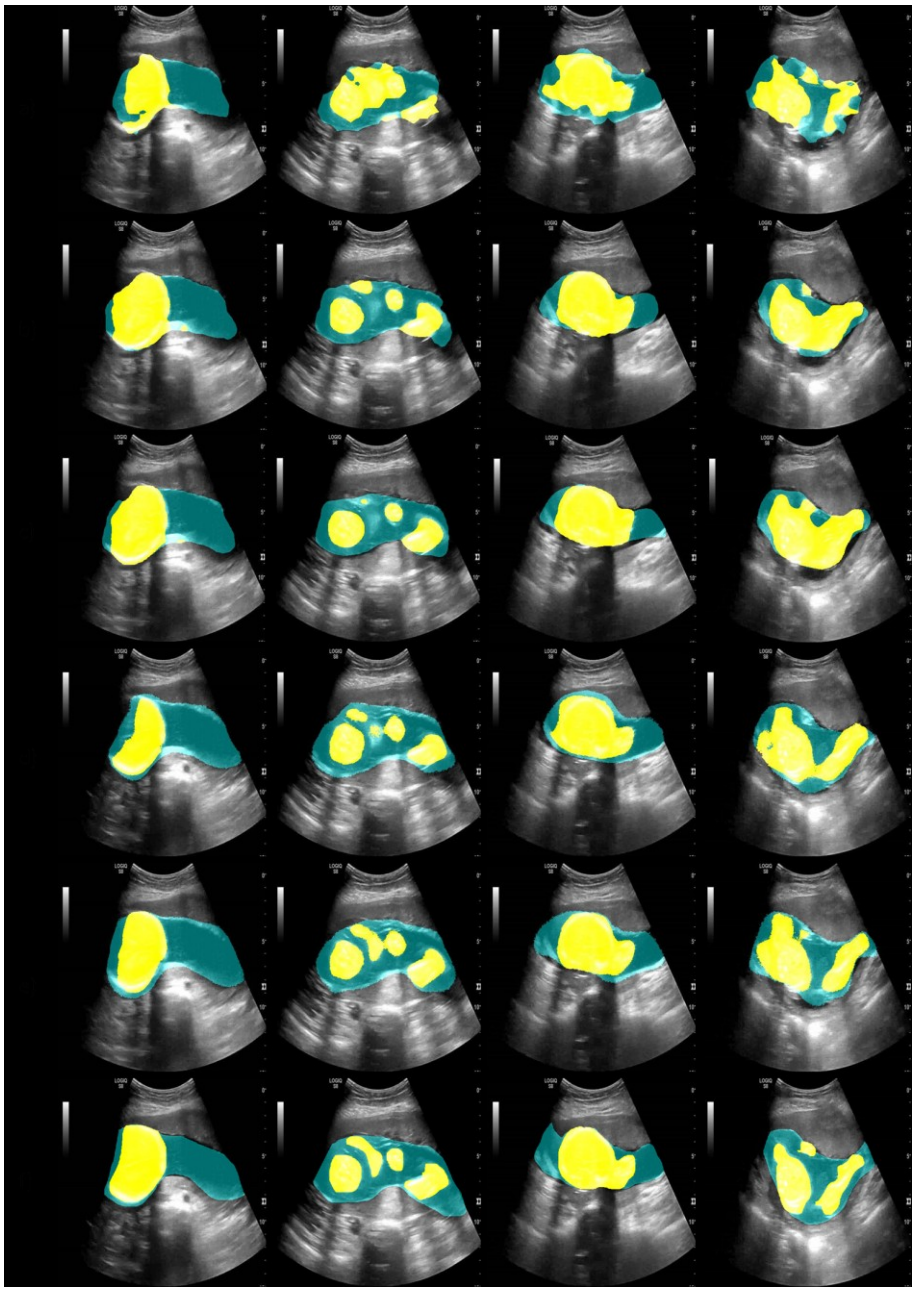

Figure 4.15 Visualized segmentation results

Visualized segmentation results of: a) FCN8S, b) PSPVGG16, c) V3+VGG16, d) EDN_IS, e) EDN_3IL_IS, f) ground truth. Each column indicates one different sample.

thesis adds 4 pooling operations after the last convolution layer of VGG16 (conv_5_3) with different size of pooling windows, then uses interpolation to up-scale the feature maps to the

Table 4.6 Evaluation results of DeeplabV3+ and PSPNet with Resnet50 as the backbone network. (%)

	IOU_fb	IOU_af	IOU_bkg.	mIOU	Accu
PSPNet	61.48	50.41	94.17	68.69	92.95
V3+_RES50	61.79	51.80	94.83	69.47	93.50

same size and concatenate them into one. The compared models with VGG16 as the backbone structure are named V3+_VGG16 and PSPVGG16. Numerical results of the above models are demonstrated in Table 4.5 and Figure 4.16. Some visualized examples of the extracted contours and segmented blobs with different approaches and the proposed methods can be found in Fig. 4.14 and Fig. 4.15.

In addition, this thesis further provides the results of Deeplabv3+ and PSPNet with Resnet50 [9] as the backbone network in Table 4.6 (V3+_RES50 and PSPNet). Interestingly, the numerical results do not achieve better scores than using VGG16's model in the experiments. It might be because this thesis uses relatively small batch size (batch size=6 for both models) on a single GPU device. Even if this thesis has trained the models with more iterations, it seems that the performances of networks trained with batch normalizations (i.e Resnet50, Resnet101) heavily depend on the large batch size.

Note that in this module this thesis does not extend the proposed optimizations to other backbone networks such as Resnet. The reason is because the scheme records geometric representation and makes use of more local position information which is obtained from max-pooling operations. However, the original Resnet does not provide such down scaling method which is required by the currently used decoding scheme. Regarding the limitation of the work, the symmetric designed segmentation scheme has drawbacks on high GPU memory cost in the training, which could cause that the training process on the currently used device might encounter GPU memory problems. It limits the module to adopt the backbone networks with less learn-able weights. Therefore, this module does not concern much about the backbone networks.

Overall, the best performance in the quantitative results is the proposed optimized

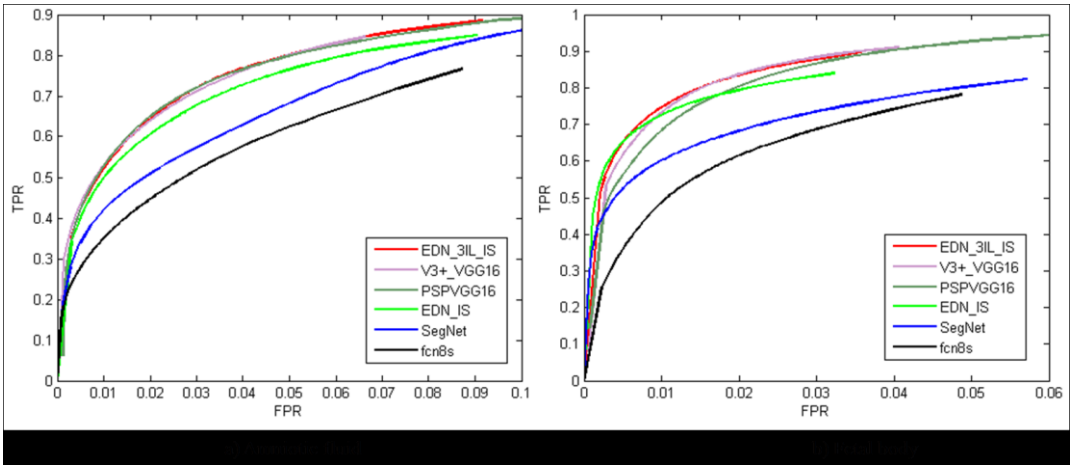

Figure 4.16 Category specified ROC curves of different models.

encoding-decoding model with 3 inner layers and intermediate supervisions. Although the result does not have large gaps with V3+_VGG16 and PSPVGG16, it ca be said that it can be proved more complete contour of objects. The proposed inner layers bring similar effects as the atrous convolution or multi-scale pooling by introducing larger field of views to the feature maps. Compared with PSPVGG16, the additional intermediate supervision branches not only have advantages in faster converging speed in the training, but they also bring smoother segmentation results by using different scales of ground truth masks in training. From the visualized results, the FCN8s model shows under fitting to the ground truth area. The proposed model performs more complete and accurate contour information on the segmentation results than others. Despite this, visualized samples show that the model with intermediate supervision have more smooth segmentations in the border areas.

On the other hand, the visualized results demonstrate that, as inner layers increase, the overall performance gets better. The network with the three inner layers achieves the best performance among all of the three indexes, and the model without inner layers performs relatively worse than others. The feature learned by additional layers is more conducive to suppress false positives in anatomical structures such as the uterus in fetal US images. On the other hand, the optimized multiple supervision module significantly improved the smoothness of the segmentation borders. With the help of the proposed multiple supervision layers, the

method does not use any post processing methods such as condition random field or morphology operations and still can achieve smooth segmentation results. In addition, the model without pre-training or data augmentations gets lower scores. The experiments show the importance of pre-training, it further demonstrates that it could bring positive effects to the cross-domain tasks by transferring learning from nature images to US images.

4.4 Conclusion

This chapter first proposes preliminary research to CNN based segmentation work for the pregnant uterus and identifies the existing issues. Then, this thesis further proposes optimized frameworks and applies the method to segmentation of multiple anatomical structures. To segment the amniotic fluid and fetal body in the uterus, this thesis adopts a fully convolution network in an encoding-decoding architecture. The input US images first are mapped into multi-channel down-scaled feature maps and then up-scaled to original size by symmetric designed decoding structures. The final segmentation results are calculated by thresholding confidence maps. It was found that the overall segmentations lack accuracy in case of some irregular shapes and not smooth enough at border areas. In order to relieve the data imbalance issue, a 2-tier approach is introduced by segmentation in the cropped uterine area. This thesis further uses stacked inner layers and intermediate supervision structure to improve the overall segmentation accuracy and smoothness at the boundaries of the segmentation results. Comparative experiments are conducted to verify the effectiveness of the proposed methods. This thesis concludes the results of this chapter as follows:

1) A fully convolutional network for semantic segmentation of pregnant US images based on backbone networks which are commonly used in nature image classification tasks. The deep learning based segmentation framework achieves pixel-wise classification in US images through hieratically decomposed feature maps and end-to-end learning using manually annotated ground truth label map.

2) A 2-tier segmentation approach which adopts the bounding box of the uterus to reduces a large number of background areas. Compared with segmentations applied directly to raw

US images, segmentation in the cropped uterine area optimize the segmentation results by relieving the imbalance issue and aligning the region of target pixels.

3) The inner layer with 1x1 sized convolution kernels. The additional convolution operations between encoding and decoding structures improve the overall segmentation accuracy by extending the network to larger dimensional space and enhanced global representations.

4) The intermediate supervision structures. The intermediate supervision of multiple down-scaled ground truth label maps brings smoother segmentation results by adding supervisions with different size of down-scaled factors for each group of output layers with different feature levels.

The experiments demonstrate the performance of the designed baseline structures and the effectiveness of the proposed optimizations. The averaged pixel-wise classification accuracy is about 93% and averaged intersection of the union is about 73%. The quantitative results of the proposed model outperform all of the other segmentation approaches. The visualized results demonstrate smooth segmentations than other methods.

Regarding the future work of the research, it can be said that in order to provide more useful semantic information for the subsequent systems, more fine-grained categories are desired to be recognized such as the fetal head, brain, and bones in the given pregnant US images.

It is found that in the proposed two-tier segmentation approach, the separate calculations of uterus detection and anatomical structure segmentation by using two independent models are computation redundant. Future study might have a chance to seek for a joint learning architecture to learn the multiple tasks (the detection and segmentation) with shared convolution weights simultaneously.

On the other hand, the limitation of extending the proposed optimizations to more advanced backbone structures such as Resnet should be dealt with. First, the structure needs to be further modified to fit the decoding structure used in the proposed segmentation scheme. Then,

the extra memory cost caused by the symmetrically designed scheme needs to be reduced by making the network more lightweight to avoid the computation limitation without losing too much performance.

Furthermore, to train the segmentation model, medical professions are asked to carefully label the anatomical structures on each of the frames. Such fully annotations on the border of each of the anatomical structures in US images are difficult and time-consuming for future medical applications.

Chapter 5. Weakly Supervised Region Mining of Fetal Head

5.1 Introduction

As mentioned in Chapter 1, automatic fetal care systems are desired by medical areas. In order to prove such high-level semantic information for post-sequence systems, this thesis has already introduced two of the proposed modules in Chapter 3 and Chapter 4, which are: bounding box localization of the uterus and semantic segmentation of anatomical structures in pregnant US images.

Next, this thesis proposes to introduce a weakly-supervised learning method for region mining of fine-grained anatomical structure. Such a system requires a technology that locates the fetal head so as to infer the gesture and position of the fetus; then, the system can perform subsequent processes such as guiding the US probe to the desired positions for further measurement. Furthermore, through the above introduced modules, it is found that the manual annotation of the pixel-wise segmentation is very time-consuming and cost relatively high. Therefore, the fetal head is desired to be separated from other fetal body parts in an efficient and low-cost manner. This chapter proposes a fetal head region mining method based on a weakly-supervised approach.

In this chapter, the major target is to propose a weakly supervised approach for localizing anatomical structures which are difficult to be annotated. Classification of the fetal head plane in US image has been studied for a long time as introduced in the previous chapters, e.g. by H. Chen et al [98]. However, in many cases of automatic medical treatments, classification of the category of US planes is not enough. The model needs to locate the position of the targets in

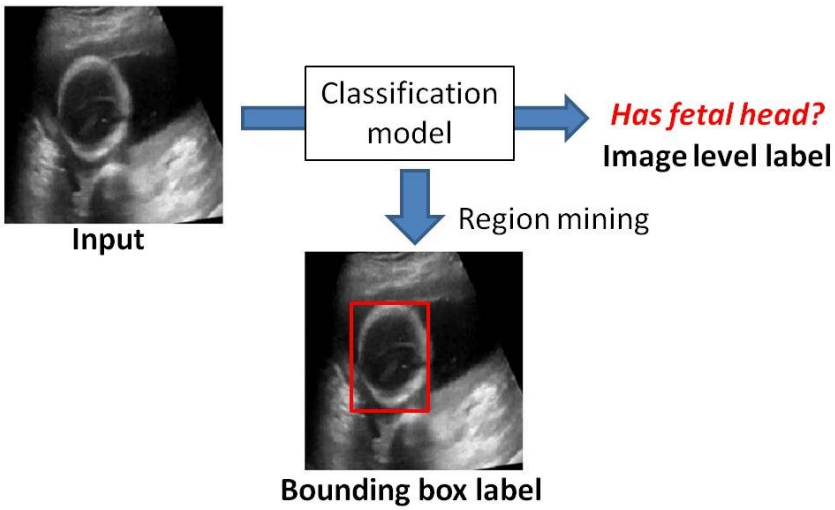

Figure 5.1 Localization of fetal head by learning from image level annotations.

the image in order to provide more reliable references. Normally, object detection or pixel-wise classification model can be obtained by fully supervised learning, as introduced in the previous chapters such as N. B. Albayrak et al. [67]. However, these methods rely on accurate manual annotation of the target position, such as the pixel-wise classification mask or bounding boxes, which cost a large number of human resources.

The hieratical feature can be used to visualize the discriminative region by the method proposed by B. –L. Zhou et al. [55]. The convolutional neural network maps the input data into hieratical feature spaces; then the feature maps can be adopted as a salient image of the target object through linear combination. In case of medical image processing, recently, some of the publications directly use the method to visualize the interesting area of the illness, for instance, X. –S. Wang et al. [99]'s work. X. –S. Wang et al. mainly target the classification tasks. The extracted response of the last convolution layers has a small size and high feature level, which causes the response map only reflects the regions which have the most discriminative information. Y. –C. Wei et al. [100] propose to solve the problem by iteratively masking the pixels which have high confidence, then by re-classifying the image from the rest of the pixels. However, the method cannot provide a reliable termination method; i.e. the mining region has a risk to be over-segmented. Therefore, the issue of the completeness of the

weakly supervised region mining still needs to be further solved.

Concerning the weakly supervised object localization approach for US images, C. F. Baumgartner et al. [54] leverage the work of [58] to visualize the discriminative areas of the fetal body in US images by back propagation, and improve the localization accuracy by using the saliency map [57] as a weighted linear combination of the back prorogation results. In their following works [59] [60], they further extend the method by introducing attention mechanism etc. Their methods suppress the noise in the generated saliency map to some extent. However, the back prorogation and global max-pooling operation are limited to a point of the most discriminative area rather than determining the full extent of the object. In other words, their methods face the issue of the extracted areas could hardly coincide with the entire object, which is very important for some medical applications such as the reconstruction of the fetal body. In addition, the proposed results lack quantitative comparison with others. N. Toussaint et al [61] propose to directly adopt the work [56] which is originally applied for nature images to extract the saliency area of the fetal body in US images. Their experiments are implemented on a notebook PC with a relatively shallow backbone network and lack of quantitative comparisons.

The targets of this model are illustrated in Figure 5.1. Given US images, through learning using image level annotations, the method aims at mining the region of the fetal head from the learned models. The region of the fetal head is represented as a tight bounding box of the fetal head area. This thesis first adopts a fetal head classification framework based on existing backbone networks and then proposes to use discriminative maps of the fetal head which are merged in multi-scale feature maps to improve the completeness of the mined region.

5.2 Region mining of fetal head from image level annotations

5.2.1 Fetal head plane classification from US images

The modified CNN models are used to train classifier in US slices and image level fetal head annotations, which are manually specified if each image includes a fetal head. To classify the input US image into different categories, the CNN model uses a hieratical designed backbone

network to extract the feature. In this thesis, author treats off-the-shelf networks as the feature extractors, and their base structures remain unchanged.

In order to visualize the discriminative area from learned classification models, the essential modifications come from the output of the network structure. In order to map the output into one-dimensional vector, which indicates whether a fetal head exists, the normal CNN classification model first needs to reshape the responses of the convolution layers into a vector and then connects the feature vectors to inner production layers. The method needs to use discriminative feature maps to localize the fetal head area. Therefore, the network structure (for extracting feature maps, which is introduced in the next section) replaces the reshape and inner product operations of the selected backbone structures by a single global average pooling layer.

Regarding the loss function, during the training phase, the method uses typical cross entropy loss to update the weights of the binary classification model. In the sample US video sequence, fetal heads exist only in a small number of the sequences. This leads to the situation in which the data distribution has severe imbalances. Therefore, this thesis modifies the loss function with weights of the negative category. The weight ω is calculated by $\omega = N_{\mathrm{Pos}}/N_{\mathrm{Neg}}$, where N_{Pos} and N_{Neg} are the number of positive and negative samples. For each batch with N samples, the loss function of fetal head classification network can be written by

$$L_{\mathrm{fh}} = -\frac{1}{n}\sum_{n=1}^{N}[y_n \log c_n + \omega(1-y_n)\log(1-c_n)], \tag{5.1}$$

where y_n, c_n indicate ground truth and predicted confidence of image n, respectively.

5.2.2 Localization of fetal head by multi-scale discriminative maps

The method proposed by Zhou et al. [55] adopts the feature maps' responses of the last convolution layer and the weights of the output layer to obtain the discriminative map M from the learned model. In particular, the input blob of the output layer is calculated from global averaging pooling (GAP) over the feature maps f of the last convolution layer. Then

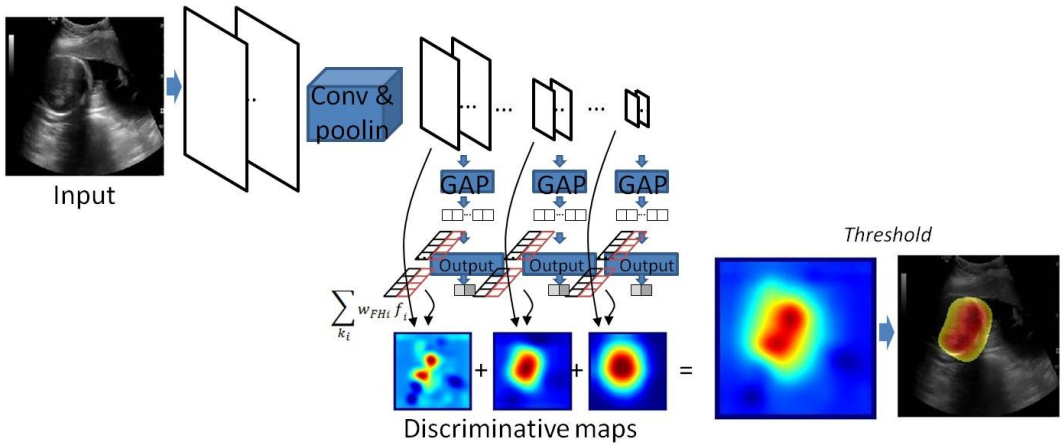

Figure 5.2 Proposed optimizations for complete fetal head region extraction. Merging of multi-scale discriminative maps.

the compressed feature is directly mapped into the shape of the output vector by linear combination with the weights w corresponding to the desired categories. The approach can be represented by,

$$M = \sum_{k} w_k f_k \tag{5.2}$$

where k is the kth channel of w and f.

The highest feature level in the deep learning model is learned from the most discriminative areas of the input feature maps.

However, the most discriminative area cannot be seen as the complete area of the target object. The mined region lacks completeness due to the presence of some of not so important areas. This finally causes the located fetal head area to be not good enough.

The pooling layers are used multiple times to narrow down the original image through averaging or maximum operations. In this study, the response maps for the discriminative localization are extracted from the last output of the hieratical layers. As the feature level becomes deeper; the semantic features get higher. The less important feature areas are progressively ignored by the network. This thesis assumes that the ignorance could degrade the completeness of the target shape. To deal with the issue, the method proposes to adopt the

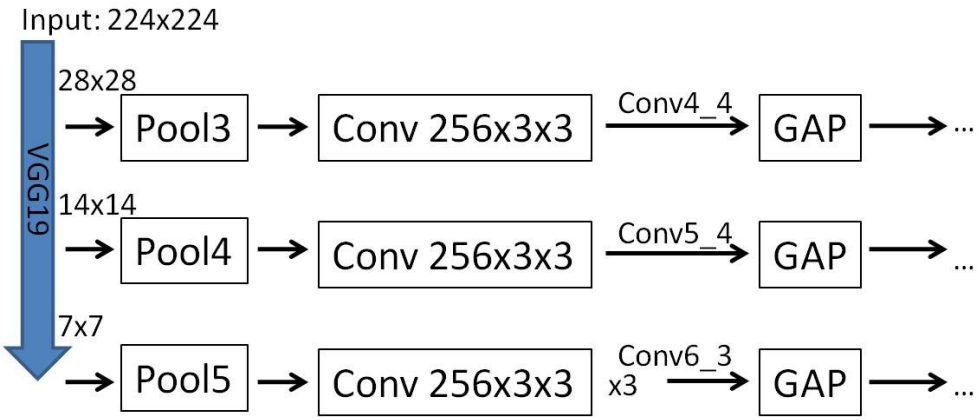

Figure 5.3 Detailed structure of output branches added on VGG19.

outputs from the multi-level features. The response maps from different feature levels represent different discriminative locations.

To deal with the above-discovered issues, in this module, the proposed method replaces the output by multiple output branches which split from a different level of backbone network's outputs and merge the multiple discriminative maps into one, as shown in Figure 5.2. In particular, for each output branch, this thesis adds a specified number of convolution layers and global average pooling operations. During the training, the output of each branch is compared with the ground truth labels, and the weights of each output branches are independently updated. By merging multiple outputs, the models learn discriminative response maps from layers with different feature levels. The branch structure with detailed parameters on VGG19 [48] structure can be found in Figure 5.3. During the localization, the method extracts the response maps of the last convolution layers from each output branch and use the weights of each branch separately to linearly combine them as multiple discriminative maps. The acquirement of the fetal head salience maps of the proposed optimized approach can be presented as

$$M = \sum_{i} \frac{1}{k_i} \sum_{k_i} w_{ki} f_{ki} , \tag{5.3}$$

where $\frac{1}{k_i}\sum_{k_i} w_{k_i} f_{ki}$ is the discriminative map which adopts the weights and feature maps

from the ith convolution layer. w_i, f_i and k_i indicate the weights that correspond to the fetal head, feature maps, and channels of the ith output branch, respectively. The method proposed in this chapter merges discriminative maps which are extracted from multiple branches and uses it as the final salient image of the fetal head.

5.2.3 Threshold

The output of the region is the heat map of the fetal head, which means that each pixel corresponds to the confidence of whether each pixel belongs to the fetal head area. Therefore, in order to convert the results to a bounding box from this kind of heat maps, the method needs to obtain the tight bound box from connecting regions larger than a given threshold. In this preliminary experiments, the discriminative map is first normalized to [0,1] by min and max values. Then, this study uses a fixed value (0.8) to obtain binary masks. The bounding box of the fetal head is obtained from the connected domain. To determine the threshold value, this thesis quantitatively evaluates the IOU of region mining results on the training sets and roughly selects the values which have the best performance on each subfolders. Note that the method only keeps the bounding box which has the largest area if multiple areas are obtained.

5.2.4 Backbone network

The weakly supervised region mining is different from classification tasks. Therefore, multiple candidate backbone networks are compared in this study. This thesis chooses three popular backbone networks: 1) VGG19 [48], 2) Alexnet [7], and 3) Resnet50 [9].The detailed parameters of each network are shown in Table 5.1, Table 5.2, and Table 5.3. The method seeks for the best structures through experiments.

5.3 Numerical Results and Discussions

5.3.1 Dataset and training details

Data collection This thesis conducts experiments on clinical US dataset to verify the validity of the proposed method of weakly-supervised region mining of the fetal head. The research acquires four clips of US pregnant examinations from a hospital as the metadata. For each US

video clip, an anonymous patient with different fetal weeks that arranged from 19 to 23 weeks is used for the experiments. The 2-folder cross-validation is adopted by alternatively training and testing on randomly selected groups of US clips (each group has two clips) with different subjects. This thesis selects the clips that have fetal body perpendicular to the US scan plane. After pruning and interval sampling operations, the US video clips are stored in a sequential image format. Cross-validation is used to alternatively train and test on each US clips. Professions are asked to manually identify the frames in which the fetal head can be observed. Then, they are asked to assign pixel-wise labels to each of the images containing fetal heads. Note that the pixel-wise annotations are only used for evaluation purpose. The bounding boxes of the fetal head are calculated from thus obtained segmentation masks.

Data augmentation The dataset has severe data imbalance and lacks diversity. Therefore, the method uses related large scaled data augmentation operations on the training set. The random crop, rotation, scale transform, and horizontal flip are added to the raw US image to give disturbances to each of the training samples. The augmented samples are resized to the fixed size in order to fit the input of the backbone network.

Domain transferred learning Compared with learning from random initialed weights, better initializations can be obtained from pre-trained low-level representations that learned from large-scale cross-domain images. Most of the weights of the convolution layers are updated from ImageNet [47] pre-trained models, while the additional layers with learn-able weights are learned from scratch.

5.3.2 Evaluation metric

First, to evaluate the model for fetal head plane classification, the experiments adopt precision, recall and F_1 score. They can be defined as

$$F_1 = 2 * \frac{\text{Recall} - \text{Precision}}{(\text{Recall} + \text{Precision})}, \tag{5.4}$$

Table 5.1 Backbone network architecture: VGG19

Conv. Layers of VGG16
Conv1_1: 64x3x3, Stride: 1, Pad: 1
Conv1_2: 64x3x3, Stride: 1, Pad: 1
Pool1: 3x3, Stride: 2
Conv2_1: 128x3x3, Stride: 1, Pad: 1
Conv2_2: 128x3x3, Stride: 1, Pad: 1
Pool2: 3x3, Stride: 2
Conv3_1: 256x3x3, Stride: 1, Pad: 1
Conv3_2: 256x3x3, Stride: 1, Pad: 1
Conv3_3: 256x3x3, Stride: 1, Pad: 1
Conv3_4: 256x3x3, Stride: 1, Pad: 1
Pool3: 3x3, Stride: 2
Conv4_1: 512x3x3, Stride: 1, Pad: 1
Conv4_2: 512x3x3, Stride: 1, Pad: 1
Conv4_3: 512x3x3, Stride: 1, Pad: 1
Conv4_4: 512x3x3, Stride: 1, Pad: 1
Pool4: 3x3, Stride: 2
Conv5_1: 512x3x3, Stride: 1, Pad: 1
Conv5_2: 512x3x3, Stride: 1, Pad: 1
Conv5_3: 512x3x3, Stride: 1, Pad: 1
Conv5_4: 512x3x3, Stride: 1, Pad: 1
Pool5: 3x3, Stride: 2

Table 5.2 Backbone network architecture: Alexnet

Conv. Layers of AlexNet
Conv1_1: 96x11x11, Stride: 4, Pad: 1
Pool1: 3x3, Stride: 2
Conv2_1: 256x5x5, Stride: 1, Pad: 1
Pool2: 3x3, Stride: 2
Conv3_1: 384x3x3, Stride: 1, Pad: 1
Conv3_2: 384x3x3, Stride: 1, Pad: 1
Conv3_3: 256x3x3, Stride: 1, Pad: 1
Pool3: 3x3, Stride: 2

where $\text{Recall} = N_{\text{TP}}/(N_{\text{TP}} + N_{\text{FN}})$, $\text{Precision} = N_{\text{TP}}/(N_{\text{TP}} + N_{\text{FP}})$, the N_{TP}, N_{FN} and N_{FP} are the numbers of true positive, false negative and false positive predictions,

Table 5.3 Backbone network architecture: ResNet50

Conv. Layers of ResNet50	
Conv: 64x7x7, Stride: 2, Pad: 3	
Pool1: 3x3, Stride: 2	
Conv: 64x1x1, Stride: 1, Pad: 0	
Conv: 64x3x3, Stride: 1, Pad: 1	
Conv: 256x1x1, Stride: 1, Pad: 0	X3
Conv: 128x1x1, Stride: 1, Pad: 0	
Conv: 128x3x3, Stride: 1, Pad: 1	
Conv: 512x1x1, Stride: 1, Pad: 0	X4
Conv: 256x1x1, Stride: 1, Pad: 0	
Conv: 256x3x3, Stride: 1, Pad: 1	
Conv: 1024x1x1, Stride: 1, Pad: 0	X6
Conv: 512x1x1, Stride: 1, Pad: 0	
Conv: 512x3x3, Stride: 1, Pad: 1	
Conv: 2048x1x1, Stride: 1, Pad: 0	X3

respectively. This study also provides the area under curve (AUC) of ROC. The definition of AUC can be seen in Figure 5.4. Regarding the localization accuracy, the method obtains the bounding box from manually annotated segmentation masks of the fetal head. The metric adopts the intersection of union (IoU) to evaluate the results. The IOU is defined as

$$\mathrm{IOU} = (S \cap S_{\mathrm{GT}})/(S \cup S_{\mathrm{GT}}), \tag{5.5}$$

where S and S_{GT} are obtained area of the predictions and the ground truth. Besides IOU, of the bounding box (Bbox_IOU), the pixel-wise IoU (Pwise_IOU) is also provided. The calculation of pixel-wise IOU is similar to the bounding box IOU, except the research judges the classification accuracy at each of the pixels of the input image.

5.3.3 Results and discussions

Regarding the classification model, the experiments compare several popular backbone networks, which are Alexnet, Resnet50 and VGG19. The detailed parameters of each structure are explained in Section 5.2.3. For both Alexnet and VGG19 structures, the models are added with batch normalization [101] operations after each convolution layer. As

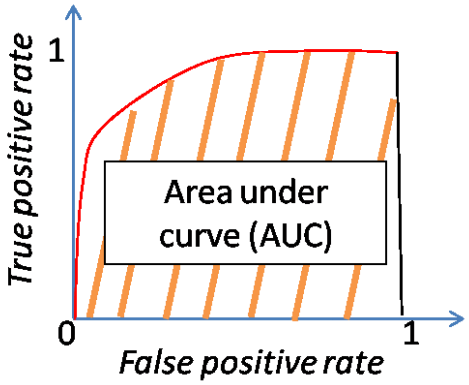

Figure 5.4 Definition of area under curve (AUC).

mentioned in Section 5.2.1, the method replaces all of the fully connecting layers with the convolution layer and global average pooling operations.

The comparison results of the three models for the fetal head plane classification task can be found in Table.5.4. As the results show, among the selected three models, the VGG19 with batch normalized structure achieves the best performance. Although the AlexnetGAP has a simple structure and the smallest network scale, it still can get acceptable classification rate because of the batch normalization and the pre-trained weights. The following experiments build optimizations based on VGG19 structure.

The comparison of the IOU of the bounding box of the three backbone structures can be seen in Table.5.5. In Table.5.5, to decouple the localization and classification performance, the scores (not in brackets) do not count false negative classification samples, while the scores in brackets are the IoU results with false negative classifications. The results of Resnet50GAP are worse than AlexnetGAP probably because it has too small feature map size. Same as the classification results, the VGG19 achieves the best localization results. By adding the proposed optimization method (VGG19GAP_OutputMerge), the localization performance is further improved. This thesis compares the result of only using the last feature map output and the two optimization methods in Table.5.6. The name of each result indicates the output of different layers in VGG19GAP. Regarding the proposed multiple discriminative outputs, the method adopts the conv4_3, conv5_3, and conv6_3 from VGG19 structure (the detail is

Table 5.4 Classification results of different backbone networks (%)

	AlexnetGAP	Resnet50GAP	VGG19SP	VGG19GAP
Recall	87.79	86.92	86.38	**90.92**
Precision	88.65	90.73	**94.94**	90.26
F1 score	87.62	88.39	90.32	**90.35**
AUC	95.87	96.34	96.76	**96.80**

Table 5.5 Localization results* with different backbone networks (%)

	AlexnetGAP	Resnet50GAP	VGG19SP	VGG19GAP	VGG19GAP _OutputMerge
Bbox_IoU	41.4 (37.0)	40.3 (36.1)	60.5 (54.9)	61.2 (54.8)	**65.3 (58.0)**
Pwise_IoU	60.7 (58.8)	57.7 (55.0)	70.3 (69.2)	72.0 (70.8)	**76.5 (73.1)**

*Values not in bracket: w/o false negative classifications. Values in bracket: w/ false negative classifications.

Table 5.6 Localization results with different output strategies (%)

	Conv4_4	Conv5_4	Conv6_3	OutputMerge
Pwise_IoU_FetalHead	19.83	51.51	54.95	**57.94**
Pwise_IoU_Bkg.	92.80	**95.40**	94.83	95.13
Bbox_IoU	53.20	62.24	55.71	**65.25**
Pwise_IoU	56.32	73.46	74.89	**76.53**

Table 5.7 Localization results of weakly and fully supervised methods (%)

Method	Weakly supervised	Fully supervised	Bbox_IoU
VGG19GAP	✓	×	61.2
VGG19GAP_OutputMerge	✓	×	**65.3**
FRCNN12)	×	✓	72.2

shown in Figure 5.3). The result obtained by merging the above-mentioned three outputs is indicated by "OutputMerge". From the quantitative results they proved that the merged feature maps bring significant improvements by enriching the integrity of the fetal head area. Some of the results and their ground truth annotations are shown in Figure 5.5. More visualized results are demonstrated in Figure 5.6. The examples in the first three rows are the true positive predictions, while the examples on the last row are wrong predictions.

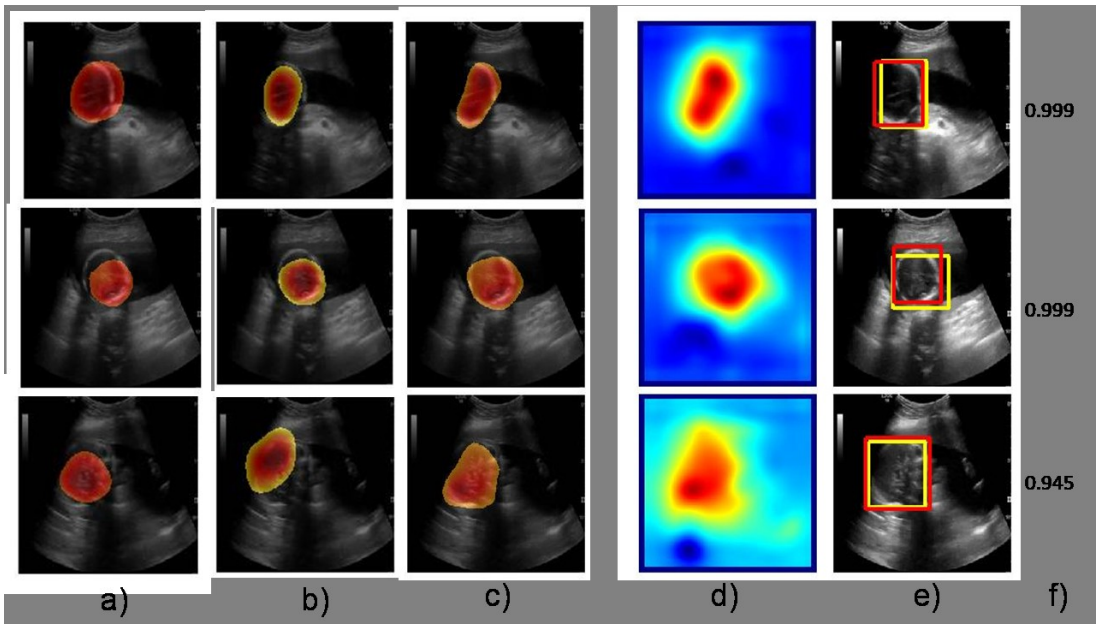

Figure 5.5 Some of the visualized results obtained by different approaches. a) VGG19GAP (CAM), b) VGG19SP, c) VGG19GAP_OutputMerge (proposed), d) merged heatmap of VGG19GAP_OutputMerge, e) obtained bounding box of VGG19GAP_OutputMerge (yellow) and ground truth (red) of fetal head region, f) classification probability of positive category.

Concerning the comparisons with related work applied in US images, this thesis provides the results of comparing with [61] and shows them in Table 5.4 and Table 5.5. Note that for fair comparison purpose, this thesis extends the backbone structure used in [61] from VGG13 to VGG19, which achieves the best performance in this research. The method is based on the network with soft proposal [56] modules; thus here the results are indicated by "VGG19SP". In the classification results (Table 5.4), the VGG19SP achieves the similar accuracy to standard VGG19GAP with higher precision and lower recall. It seems that the effect of the backbone network plays a dominant role as classification task. On the other hand, the localization results (in Table 5.5) shows that the model with the proposed optimizations achieve better performance than VGG19SP. Some of the visualized results are demonstrated in Figure 5.5.

In addition, to evaluate the accuracy difference with the fully supervised approach, this thesis further demonstrates the results which are obtained from Faster-RCNN [10] (FRCNN). Note

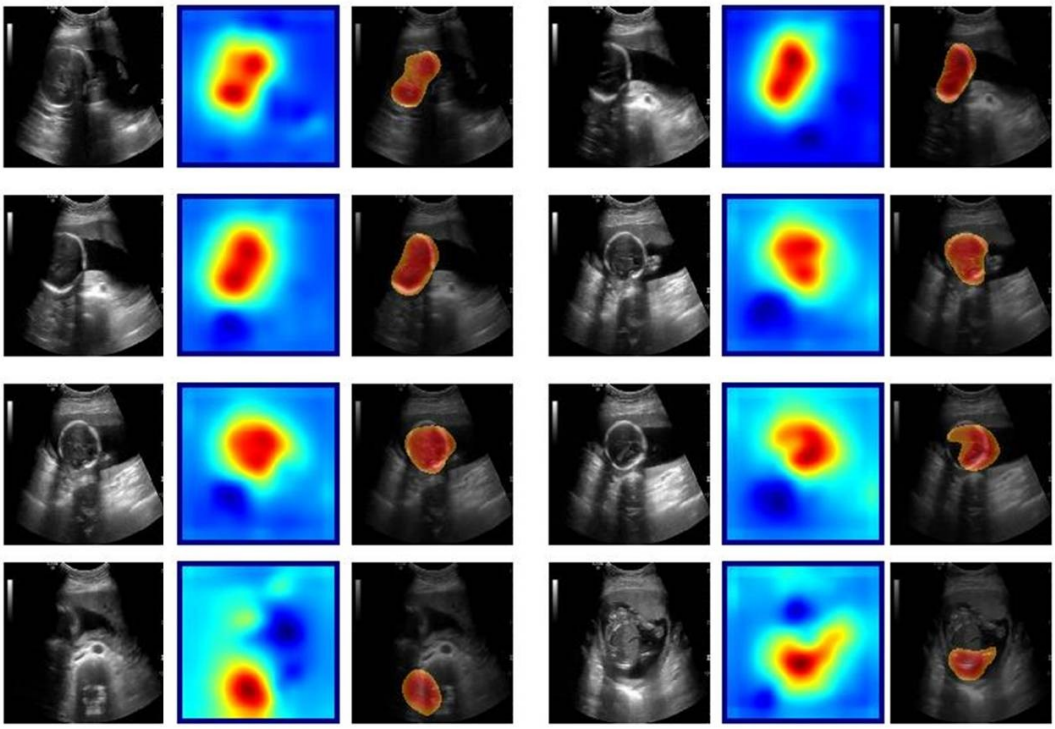

Figure 5.6 More visualized results obtained by VGG19GAP_OutputMerge.
The first and the fourth columns show the US image samples. The second and the fifth columns show the merged feature maps obtained by the proposed methods. The third and the sixth columns show the mined regions of fetal head.

that the model of FRCNN is learned from bounding box annotations. Compared with weakly supervised methods, it is impossible to learn from image level annotations. To train this model, the experiment adopts the manually labeled bounding box annotations, which are originally used for evaluation. In particular, to the structure of FRCNN model this thesis embeds the original VGG16 as the backbone structure and use 600x600 pixels as the input resolution for both training and testing. The comparison results are shown in Table.5.7. The results show that compared with the baseline weakly supervised method (first row), the method with the proposed optimizations (second row) shorten the accuracy gaps between the models which learn from image level annotations and from object position annotations. On the other hand, the weakly supervised method has strong advantages over learning only from image level annotations, in that the weakly supervised method can avoid complex and time-consuming medical image annotation tasks

5.4 Conclusion

This chapter proposes a weakly supervised method to obtain the fetal head area in US images. The method first adopts modified CNN architecture as a binary classifier to learn if the input US plain contains a fetal head or not. Then the feature maps of the model can be used to visualize the fetal head area as the most discriminative region which can distinguish the positive and negative samples. The learned weights of the last output layer are used to sum up the feature maps.

Through experiments, the study has findings and discussions about the insufficiency of the current method. The currently used class activation mapping method lacks completeness in the entire object region. In response, the research finds that the feature maps of multiple feature levels represent a different discriminative region. Therefore, this thesis proposes to make use of more redundant information of multiple feature levels from different layers of the network. Multiple feature maps are extracted and merged to improve the results of weakly-supervised region mining for fetal head area.

As the main contribution of this module, this thesis inspires the future research to merge multi-scaled discriminative maps with different feature levels to get more complete salient areas. The effectiveness of the proposed method is verified through preliminary experiments. As the results, the method achieves higher than 95% AUC for fetal head classification and 76% for overlapping accuracy in IOU with manually labeled ground truth, which outperforms the baseline deep learning approach and other related works.

Chapter 6. Conclusion

6.1 Summary of Thesis

To extract accurate high level semantic information of important tissues for raw US images. In this research, the author explores the feasibility of deep learning techniques in the US image processing area. In particular, this thesis has proposed three related modules that localize and extract the areas of multiple anatomical structures in US images by end-to-end learned CNN based architectures and weakly supervised learning. The proposed solutions discuss the limitation of existing tradition or deep learning methods. To verify the effectiveness proposed optimizations and compare with previous works, the proposed algorithms are evaluated through quantitative experiments and intuitive visualized results on pregnant US image slices obtained by clinical examinations.

In summary, the three modules that are proposed in this thesis are 1) uterus localization using a bounding box regression CNN, 2) semantic segmentation of multiple anatomical structures using CNN and its optimizations, and 3) weakly-supervised method for region mining of the fetal head. From coarse to fine, the methods 1), 2) and 3) are explained in Chapter 3, Chapter 4 and Chapter 5, respectively. The conclusions of each module that is proposed in this thesis are summarized as follows.

- Localization of uterus

A specifically designed CNN regression network is used to localize candidate positions of the bounding box of the uterus in raw US images. The input image is mapped to a vector with a specific length that corresponds to a set of the offset and probability values through stacked dimension transformation operations. The dimension transformation contains convolution,

pooling, non-linear activation, and inner production operations. The (output) offset values are used to transfer multiple densely designed reference boxes to target positions with high probabilities of containing the uterus, and the (output) probability values are used to judge if the corresponded reference box belongs to a positive object (uterus). The final position of the uterus is obtained by eliminating redundant candidates by non-maximum suppression algorithm. The design of the backbone network follows the existing work in the nature image processing area. The CNN regression model is trained in an end-to-end manner through the manually annotated position of the uterus.

Experiments are conducted through an evolution of the clinical dataset. In the best setting of the trained model, the intersection of union between the predicted bounding box and ground truth achieves 62%, which is higher than other related deep learning based approaches, which are originally applied for nature image domains. What is more important, the localized uterus area helps to improve the subsequent semantic segmentation by suppressing the data imbalance issue.

- Semantic segmentation of the anatomical structure

The method follows an encoder-decoder CNN architecture as a baseline to provide pixel-wise classifications in US images to segment the areas of desired anatomical structures, which are a uterus, amniotic fluid, and fetal body. The model first maps the input image into down-scaled feature maps through sets of convolution, pooling, and non-linear activation operations. Then, in order to achieve pixel-wise segmentation, the model transfers the feature maps back to the same sized input image and outputs a feature matrix in the specific format through recorded maximum indexes used in pooling operations. The methods proposed in this thesis further improve the performance in terms of several aspects. To enhance the global shape information to increase the segmentation accuracy, multiple inner layers with 1x1 sized convolution kernels are inserted between encoding and decoding parts. To improve the smoothness of the segmented blobs, the network structures with intermediated supervision operations are proposed. In addition, as mentioned earlier, the method makes use of localized

uterus area to relieve the imbalance issue.

The methods are evaluated on cleaned US images through cross-validations. The averaged pixel-wise classification accuracy is about 93% and averaged intersection of union is about 73%. The visualized results demonstrate smoother segmentations than other deep learning based methods compared. The acceptable results prove that the proposed methods can be used for subsequent automatic systems.

- Weakly-supervised region mining of fetal head

A specifically designed CNN classification model and optimized region mining method are proposed to extract the complete discriminative area of fetal head from US images. The model learns to classify the plane with the fetal head by image level annotations. During testing, the image is resized and input to the classification model. Then, the probability of the fatal head slice is predicted through backbone network structures. Regarding the region of the fetal head, the cumulative activation mapping is used to extract the discriminative area of the input image. The discriminative area can be seen as the region of the fetal head. The final position of the fetal head is obtained through thresholding the merged feature maps. The original method lacks completeness on the fetal head in US images; therefore, the proposed method proposes to optimize the completeness through extracting multiple feature maps from feature levels in different size of receptive fields.

Experiments demonstrate higher than 96% classification results in AUC, and 76% overlapping accuracy in the intersection of the union. The results prove that in the pregnant US image the optimized methods can achieve good performance that outperforms original proposals and other deep learning based method.

In summary, this research has introduced deep learning based frameworks and optimizations for accurate high level semantic information extraction from various aspects. To verify the gaps between the proposed methods and real-world usages, the performance of each module is compared with human doctors with years of experience. The results are promising that this research makes the development of automatic antennal examinations one step closer to the

real world solutions. The proposed methods have potential to be used to assist the human doctors in clinical examinations or to be used as the inputs of other upper level medical application systems such as automatic amniotic fluid detection to relieve the issues of shortage of manpower in hospitals.

6.2 Future works

Remaining issues and possible solutions for future works are summarized in following.

- Multi-task learning

This thesis proposes various approaches by using multiple independent CNN models. Each model is used to solve one specific task in the US image processing. Recently, some works [102] have proved that the weights of the CNN model can be used to learn share-able feature representations from multiple tasks (or even cross-domain tasks) by using a single model. This feature allows the model to handle multiple tasks with less memory storage and faster inference speed. What is more, for some of the tasks, multi-task learning can achieve higher performances by correlations features. It is possible to design an "all-in-one" model which can simultaneously output multiple objectives such as localization of uterus and semantic segmentation of amniotic fluid and fetal body with shared network structures and weights.

- Segmentation of fine-grained structures

This thesis targets some representative organs in the antenatal examination. More structures with fine-grained categories still need to be handled to achieve completeness information for automatic systems: for example, the body parts of the fetus. The body parts of the fetus can provide more information about the pregnancy to the tasks such as gesture recognition and localization of interior robot. To distinguish the difference between multiple fetal body parts, the learning of intra class distance between different object needs to be enhanced.

- Data generation

For deep learning based techniques, data collection and annotation are always treated as the most important and difficult mission all the time, especially for medical image processing

area. Under the conditions of policy restrictions, the best way is to make more contact with hospitals and patients to ask for the permissions of data usage and profession annotations. It is no doubt that the gathering of real-world dataset will cost lots of times and risks for the researchers. Another option is the use of simulated data generated by simulation models. The demerit of simulated data is obvious: the model learned from fake data cannot be suitable to the real-world solutions very well. According to the recent development, a possible way of learning from low shot dataset is data augmentation through generative adversarial networks (GAN) [103] and its verities such as [104] [105], etc. The specifically designed CNN structure uses the adversarial scheme to supervise the generation model to generate images which have a close appearance feature with real images. The method could generate realistic images with good visual appearance and coherent feature distribution. Through joint learning with the target task, it can improve the accuracy of models by generating more images to relieve the poor performance caused by too few training dataset or bad distributions.

In summary, deep learning techniques provide large chances to bring machine learning algorithms into real-world solutions. Preliminary experiments have verified the possibility and correctness on related medical image processing tasks. In order to further improve the efficiency of doctors, future study should seize the opportunity to extend the experiments on larger scaled datasets and integrate deep learning algorithms to medical imaging devices.

Bibliography

[1] F.P. Hadlock, R.B. Harrist, R.S. Sharman, R. L. Deter and S. K. Park: "Estimation of Fetal Weight with The Use of Head, Body, and Femur Measurements—A Prospective Study", American Journal of Obstetrics & Gynecology, Vol. 151 No.3, pp. 333-337 (1985).

[2] J. F. Randolph, Y. K. Ying, D. B. Maier, C. L. Schmidt and D.H. Riddick: "Comparison of Real-Time Ultrasonography, Hysterosalpingography, and Laparoscopy/Hysteroscopy in The Evaluation of Uterine Abnormalities and Tubal Patency", Fertility & Sterility, Vol. 46 No. 5 pp.828-32 (1986).

[3] J. –J. Huang and D. –W. Hong: "The Factors Influencing Antenatal Examination of Pregnant Women and the Countermeasures", Journal of Nursing (China), No. 2 (2006).

[4] R. L. Brent: "The Effect Of Embryonic and Fetal Exposure to X-Ray, Microwaves, and Ultrasound", Clinics in Perinatology, Vol. 13, No.3, pp.615-648 (1986).

[5] U. M. Reddy, R. A. Filly and J. A. Copel: "Prenatal Imaging: Ultrasonography and Magnetic Resonance Imaging", Obstetrics & Gynecology, Vol. 112, No. 1, pp.145-157 (2008).

[6] G. E. Hinton, S. Osindero and Y. W. Teh: "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, Vol. 18, No. 7, pp.1527-1554 (2006).

[7] A. Krizhevsky I. Sutskever and G. Hinton: "ImageNet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems 25, pp.1097-1105 (2012).

[8] C. Szegedy, W. Liu, Y. –Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich: "Going Deeper with Convolutions", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp.1-9. (2015).

[9] K. –M. He, X. Zhang, S. –Q. Ren and J. Sun: "Deep Residual Learning for Image Recognition", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778 (2016).

[10] S. –Q. Ren, K. –M, He, R. Girshick and J. Sun: "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks", Proc. of International Conference on Neural Information Processing Systems, Vol.1, pp.91-99 (2015).

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. –Y. Ful and A. C. Berg: "SSD: Single Shot MultiBox Detector", Proc. of European Conference on Computer Vision, Vol.9905, pp.21-37 (2016).

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi: "You Only Look Once: Unified, Real-Time Object Detection", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp.779-788 (2016)

[13] E. Shelhamer, J. Long, T. Darrell: "Fully Convolutional Networks for Semantic Segmentation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.39, No.4, pp.640-651 (2017).

[14] V. Badrinarayanan, A. Kendall, R. Cipolla: "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.39, No.12, pp.2481-2495 (2017).

[15] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz and K. Shpanskaya: "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning", arXiv:1711.05225 (2017).

[16] "Blue Histology - Female Reproductive System Archived 2007-02-21 at the Wayback Machine", School of Anatomy and Human Biology, The University of Western Australia Accessed 20061228 20:35, http://www.lab.anhb.uwa.edu.au/mb140/CorePages/FemaleRepro/FemaleRepro.htm

[17] "Diagnostic Tests – Amniocentesis", Harvard Medical School. Retrieved 2008-07-15. http://www.health.harvard.edu/diagnostic-tests/amniosentesis.htm

[18] J. Dodgson, J. Martin, J. Boswell, H. B. Goodall and R. Smith: "Probable Amniotic Fluid Embolism Precipitated by Amniocentesis and Treated by Exchange Transfusion", British Medical Journal (Clinical research ed.), Vol. 294, No. 6583, pp. 1322–1323 (1987).

[19] Y. D. Yuan and G. L. Foley: "Female Reproductive System, A Review of Histogenesis/Organogenesis in the Developing North American Opossum (Didelphis virginiana)", Springer Berlin Heidelberg, pp. 915–946 (1998).

[20] K. Z. Abd-Elmoniem, A. B. M. Youssef and Y. M. Kadah: "Real-Time Speckle Reduction and Coherence Enhancement in Ultrasound Imaging via Nonlinear Anisotropic Diffusion", IEEE

Trans. on bio-medical engineering, Vol. 49, No.9, pp.997-1014 (2002).

[21] M. A. Underwood, W. M. Gilbert and M. P. Sherman: “Amniotic Fluid: Not Just Fetal Urine Anymore”, Journal of Perinatology Official Journal of the California Perinatal Association, Vol. 25, No.5, pp.341-348 (2005).

[22] P. C. Macdonald, S. Koga and M. L. Casey: “Decidual Activation in Parturition: Examination of Amniotic Fluid for Mediators of The Inflammatory Respons”, Annals of the New York Academy of Sciences, Vol.622, No.1, pp.315–330 (1991).

[23] F. P. Hadlock, R. B. Harrist, R. S. Sharman, R. L. Deter and S. K. Park: “Estimation of Fetal Weight with The Use of Head, Body, and Femur Measurements—A Prospective Study”, American Journal of Obstetrics & Gynecology, Vol. 151, No.3, pp.333-337 (1985).

[24] G. Pilu, M. Segata, T. Ghi, A. Carletti, A. Perolo, D. Santini, P. Bonasoni, G. Tani and N. Rizzo: “Diagnosis of Midline Anomalies of The Fetal Brain with The Three-Dimensional Median View”, Ultrasound Obstet Gynecol, Vol.28, No.4, pp.522-529 (2006).

[25] S. Yu and K. -K. Tan: “Classification of Lumbar Ultrasound Images with Machine Learning”, Asia-Pacific Conference on Simulated Evolution and Learning. pp. 287-298 (2014).

[26] X. Liu, D. Padfield and K. Krishnan: “Learning-Based Scan Plane Identification from Fetal Head Ultrasound Images”, Proc. of SPIE The International Society for Optical Engineering, pp.8320-8329 (2012).

[27] M. Marsousi, K. N. Plataniotis and S. Stergiopoulos: “Shape-based Kidney Detection and Segmentation in Three-dimensional Abdominal Ultrasound Images”, Proc. of IEEE Engineering in Medicine and Biology Society, pp. 2890-2894 (2014).

[28] B. Rahmatullah, A. T. Papageorghiou and J. A. Noble: “Integration of Local and Global Features for Anatomical Object Detection in Ultrasound”, Medical Image Computing and Computer-Assisted Intervention, pp. 402-409 (2012).

[29] G. Pons, R. Marti, S. Ganau, M. Sentis and J. Marti: “Feasibility Study of Lesion Detection Using Deformable Part Models in Breast Ultrasound Images”, Iberian Conference on Pattern Recognition and Image Analysis. Springer Berlin Heidelberg, pp. 269-276 (2013).

[30] B. Rahmatullah and J. A. Noble: “Anatomical Object Detection in Fetal Ultrasound: Computer-Expert Agreements”, Biomedical Informatics and Technology. Springer Berlin Heidelberg, pp. 207-218 (2014).

[31] R. Bharath and P. Rajalakshmi: “Fast Region of Interest Detection for Fetal Genital Organs in B-Mode Ultrasound Images”, Proc. of IEEE conference Biosignals and Biorobotics, pp. 1-5

(2014).

[32] W. Mahmud, R. Izaham and E. Supriyanto: “Boundary Detection of Kidney Ultrasound Image Based on Vector Graphic Approach”, Modern Computer, No. 3 (2016).

[33] J. R. R. Uijlings, V. D. Sande, K. E. A. Gevers and A. W. M, Smeulders: “Selective Search for Object Recognition”, International Journal of Computer Vision, Vol.104, No.2, pp.154-171 (2013).

[34] P. R. Thangaraj and P. Tamilselvi: “Segmentation of Calculi from Ultrasound Kidney Images by Region Indictor with Contour Segmentation Method”, Global Journal of Computer Science and Technology, (2012).

[35] P. R. Tamilselvi and P. Thangaraj: “Modified Watershed Segmentation Method to Segment Renal Calculi in Ultrasound Kidney Images”, International Journal of Intelligent Information Technologies, Vol. 8, No.6, pp.46-61 (2012).

[36] P. R. Tamilselvi: “Segmentation of Renal Calculi in Ultrasound Kidney Images Using Modified Watershed Method”, Recent Advances in Intelligent Technologies & Information Systems, (2015).

[37] I. Sobel and G. Feldman: “A 3x3 Isotropic Gradient Operator for Image Processing”, Die Pharmazie, Vol. 7, No. 8 (1968).

[38] H. P. Moravec: “Obstacle Avoidance and Navigation in The Real World by A Seeing Robot Rover”. Stanford University, Doctoral Dissertation (1980).

[39] P. S. Hiremath and J. R. Tegnoor: “Automatic Detection of Follicles in Ultrasound Images of Ovaries using Edge Based Method”, International Journal of Computer Applications Special Issue on Recent Trends in Image Processing & Pattern Recognition (RTIPPR), No.3, pp.120-125 (2011).

[40] J. Canny: “A Computational Approach To Edge Detection”, IEEE Trans. on Pattern Analysis and Machine Intelligence, No. 8, pp.679–714 (1986).

[41] J. Shan, H. –D. Cheng and Y. –A. Wang: “Novel Segmentation Method for Breast Ultrasound Images Based On Neutrosophic l-Means Clustering”, Medical Physics, Vol. 39, No.9, pp.5669 (2012).

[42] X. Xian, H. D. Cheng Y. Zhang: “A Fully Automatic Breast Ultrasound Image Segmentation Approach Based on Neutro-Connectedness”, Proc. of IEEE International Conference on Pattern Recognition, pp. 2495-2500 (2014).

[43] G. Slabaugh, G. Unal, M. Wels, T. Fang and B. Rao: "Statistical Region-Based Segmentation of Ultrasound Images", Ultrasound in Medicine & Biology, Vol. 35, No. 5, pp.781 (2009).

[44] G. Slabaugh, G. Unal, T. Fang and M. Wels: "Ultrasound-Specific Segmentation via Decorrelation and Statistical Region-Based Active Contours", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 45-53 (2006).

[45] W. –M. Wang, Z. Lei, Q. Jing, Y. –P. Chui, N. –L. Bing and P. A. Heng: "Multiscale Geodesic Active Contours for Ultrasound Image Segmentation Using Speckle Reducing Anisotropic Diffusion", Optics & Lasers in Engineering, Vol. 54, No. 1, pp.105-116 (2014).

[46] L. Bo, H. -D. Cheng, J. –H. Huang, J. –W. Tian, X. –L Tang and J. –F. Liu: "Probability density difference-based active contour for ultrasound image segmentation", Pattern Recognition, Vol. 43, No. 6, pp. 2028-2042 (2010).

[47] J. Deng, W. Dong, R. Socher, L. –J. Li, K. Li, F. –F. Li: "ImageNet: A Large-scale Hierarchical Image Database", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp.248-255 (2009).

[48] K. Simonyan, A. Zisserman: "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv:1409.1556, (2014).

[49] J. –F. Dai, L. Yi, K. –M. He and J. Sun: "R-FCN: Object Detection via Region-based Fully Convolutional Networks", arXiv:1605.06409, (2016).

[50] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun: "Dermatologist-level classification of skin cancer with deep neural networks", Nature, Vol. 542, No. 7639, pp. 115-118 (2017).

[51] G. M. Van, B. Ginneken, C. Hoyng, T. Theelen and C. Sanchez: "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images[J]. IEEE Trans. on Medical Imaging, Vol. 35, No.5, pp. 1273-1284 (2016).

[52] G. Carneiro, J. C. Nascimento, A. Freitas: "The Segmentation of The Left Ventricle of The Heart from Ultrasound Data Using Deep Learning Architectures and Derivative-Based Search Methods", IEEE Trans. on Image Processing, Vol.21, No.3, pp.968-982 (2012).

[53] H. Chen, Y. Zheng, J.H. Park, P. -A. Heng, S. –K. Zhou: "Iterative Multi-domain Regularized Deep Learning for Anatomical Structure Detection and Segmentation from Ultrasound Images", Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol. 9901, pp. 487-495 (2016)

[54] C. F. Baumgartner, K. Kamnitsas, J. Matthew, T. P. Fletcher, S. Smith, L. M. Koch, B. Kainz,

and D. Rueckert: “SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound”, IEEE Trans. on Medical Imaging, No. 99, pp. 1-1 (2017).

[55] B. -L. Zhou, B. , A. Khosla, A. Lapedriza, A. Oliva, A. Torralba: “Learning Deep Features for Discriminative Localization”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp.2921-2929 (2016).

[56] Z. Yi, Y. –Z. Zhou, Q. –X. Ye, Q. Qiang and J. –B. Jiao: “Soft Proposal Networks for Weakly Supervised Object Localization”, arXiv:1709.01829 (2017).

[57] M. Oquab, L. Bottou, I. Laptev and J. Sivic: “Is object localization for free? - Weakly-supervised learning with convolutional neural networks”, Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 685-694 (2015).

[58] S. Karen, V. Andrea and Z. Andrew: “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, arXiv:1312.6034 (2013)

[59] J. Schlemper, O. Oktay, C. Liang, J. Matthew, C. Knight, B. Kainz, B. Glocker, Ben and D. Rueckert: “Attention-Gated Networks for Improving Ultrasound Scan Plane Detection”, arXiv:1804.05338 (2018).

[60] J. Schlemper, O. Oktay, C. Liang, J. Matthew, C. Knight, B. Kainz, B. Glocker, Ben and D. Rueckert: “Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images”, arXiv:1808.08114 (2018)

[61] N. Toussaint, B. Khanal, M. Sinclair, A. Gomez, E. Skelton, J. Matthew and J. Schnabel: “Weakly Supervised Localisation for Fetal Ultrasound Images”, arXiv:1808.00793 (2018).

[62] Y. Li, R. Xu, A. Krohn-grimberghe, J. Ohya, H. Iwata: “Deep Learning Based Uterus Localization and Anatomical Structure Segmentation on Fetal Ultrasound Image”, Trans. on Institute of Image Electronics Engineers of Japan, (June 2019, to appear)

[63] Y. Li, R. Xu, J. Ohya, H. Iwata: “Pregnant Uterine Ultrasound Image Segmentation by Encoding-Decoding Convolutional Neural Network”, Proc. of The 5th IIEEJ International Conference on Image Electronics and Visual Computing, (2017)

[64] Y. Li, R. Xu, A. Krohn-grimberghe, J. Ohya, H. Iwata: “Automatic Fetal Body and Amniotic Fluid Segmentation from Fetal Ultrasound Images by Encoder-Decoder Network with Inner Layers”, Proc. of IEEE Conference on Engineering in Medicine and Biology Society, pp. 1485-1488, (2017)

[65] Y. Li, R. Xu, A. Krohn-grimberghe, J. Ohya, H. Iwata: “Region Mining of Fetal Head in Ultrasound Image Based on Weakly Supervised Annotations and Deep Learning”, Short

paper on Institute of Image Electronics Engineers of Japan, (2019, to appear).

[66] L. Fan, P. Santago, W. Riley and D. M. Herrington: "An Adaptive Template-Matching Method and Its Application to The Boundary Detection of Brachial Artery Ultrasound Scans", Ultrasound in Medicine & Biology, Vol. 27, No.3, pp. 399-408 (2001).

[67] N. B. Albayrak, O. Betul, A. Ayse and S. Yusuf: "Prostate Detection from Abdominal Ultrasound Images: A Part Based Approach," Proc. of IEEE International Conference on Image Processing. pp.1955-1959 (2015).

[68] L. Breiman: "Stacked Regressions", Machine Learning, Vol. 24, No. 1, pp.49-64 (1996).

[69] P. Felzenszwalb, D. Mcallester and D. A. Ramanan: "Discriminatively Trained, Multiscale, Deformable Part Model", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-8 (2008).

[70] J. S. R. Jang: "ANFIS: Adaptive-Network-Based Fuzzy Inference System", IEEE Trans. on Systems,Man,and Cybernetics, Vol. 23, No.3, pp.665-685 (1993).

[71] P. T. Akkasalgar and S. S. Karakalmani: "Abnormality Detection in Kidney Ultrasound Imaging", International Journal of Electronic Commerce Studies

[72] N. Archip, R. Rohling, P. Cooperberg and H. Tahmasebpour: "Ultrasound Image Segmentation Using Spectral Clustering", Ultrasound in Medicine & Biology, Vol. 31, No. 11, pp. 1485-1497 (2005).

[73] T. Yun and H. Shu: "Ultrasound Image Segmentation by Spectral Clustering Algorithm Based on The Curvelet and GLCM Features", Proc. of IEEE International Conference on Electrical and Control Engineering, pp. 920-923 (2011).

[74] R. A. Mukaddim, J. Shan, I. E. Kabir, A. S. Ashik, R. Abid, Z. –N. Yan, D. N. Metaxas, B. S. Garra, K. K. Islam and S. K. Alam: "A Novel and Robust Automatic Seed Point Selection Method for Breast Ultrasound Images", Proc. of IEEE International Conference on Medical Engineering, Health Informatics and Technology, pp. 1-5 (2017).

[75] M. Kass, A. Witkin and D. Terzopoulos: "Snakes: Active Contour Models", International Journal of Computer Vision, Vol. 1, No. 4, pp.321-331 (1988).

[76] Y. L. Lecun, L. Bottou, Y. Bengio and P. Haffner: "Gradient-Based Learning Applied to Document Recognition". Proc. of IEEE, Vol. 86, No. 11, pp.2278-2324 (1998).

[77] R. Girshick: "Fast R-CNN", Computer Science, (2015).

[78] R. Girshick, J. Donahue, T. Darrell and J. Malik: "Region-Based Convolutional Networks for

Accurate Object Detection and Segmentation", IEEE Trans. on Pattern Analysis & Machine Intelligence, Vol. 38, No. 1, pp. 142-158 (2016).

[79] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna: "Rethinking the Inception Architecture for Computer Vision", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826 (2016).

[80] G. Huang, Z. Liu, V. D. M. Laurens and K. Q. Weinberger: "Densely Connected Convolutional Networks", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261-2269 (2017).

[81] M. Cicero, A. Bilbily, E. Colak, T. Dowdell, B. Gray, K. Perampaladas and J. Barfett: "Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs", Investigative Radiology, Vol. 52, No. 5, pp. 281 (2017).

[82] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. Mok, L. Shi and P. A. Heng: "Automatic Detection of Cerebral Microbleeds from MR Images via 3D Convolutional Neural Networks", IEEE Trans. on Medical Imaging, Vol. 35, No. 5, pp.1182-1195 (2016).

[83] J. M. Wolterink, T. Leiner, M. A. Viergever and I. Isgum: "Automatic Coronary Calcium Scoring in Cardiac CT Angiography Using Convolutional Neural Networks", Proc. of International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer-Verlag New York, Inc. pp. 589-596 (2015).

[84] S. B. Lo, S. A. Lou, J. S. Lin, M. T. Freedman, M. V. Chien and S. K. Mun: "Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection", IEEE Trans. on Medical Imaging, Vol. 14, No. 4, pp. 711-718 (1995).

[85] G. Hinton: "Deep Belief Nets", Springer US, (2011).

[86] G. Carneiro and J. C. Nascimento: "Combining Multiple Dynamic Models and Deep Learning Architectures for Tracking the Left Ventricle Endocardium in Ultrasound Data", IEEE Trans. on Pattern Analysis & Machine Intelligence, Vol. 35, No. 11, pp. 2592-2607 (2013).

[87] O. Ronneberger, P. Fischer and T. Brox: "U-Net: Convolutional Networks for Biomedical Image Segmentation", Vol. 9351, pp. 234-241 (2015).

[88] A. Neubeck and J. V. L. Gool: "Efficient Non-Maximum Suppression", Proc. of IEEE Conference on International Conference on Pattern Recognition, pp. 850-855 (2006).

[89] Y. –Q. Jia, E. Shelhamer, J. Donahue, S. Karayev and T. Darrell: "Caffe: Convolutional

architecture for fast feature embedding", arXiv:1408.5093 (2014).

[90] N. Martins, S. M. Saad, D. Veiga, M. Ferreira and M. Coimbra: "Segmentation of The Metacarpus and Phalange in Musculoskeletal Ultrasound Images Using Local Active Contours", Proc. of IEEE Conference on Engineering in Medicine and Biology Society, pp.4097-4100 (2016).

[91] B. Georgescu, S. –Z. Xiang, D. Comaniciu and A. Gupta: "Database Guided Segmentation of Anatomical Structures with Complex Appearance", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 429–436 (2005).

[92] H. Noh, S. Hong and B. Han: "Learning Deconvolution Network for Semantic Segmentation", Proc. of IEEE International Conference on Computer Vision, pp. 1520-1528 (2015).

[93] D. M. Zeiler and R. Fergus: "Visualizing and Understanding Convolutional Networks", arXiv:1311.2901 (2013).

[94] X. Glorot, A. Bordes and Y. Bengio: "Deep Sparse Rectifier Neural Networks", Proc. of International Conference on Artificial Intelligence and Statistics, pp. 315-323 (2011).

[95] L. –C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam: "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", arXiv:1802.02611 (2018)

[96] H. –S. Zhao, J. –P. Shi, X. –J. Qi, X. –G. Wang and J. –Y. Jia: "Pyramid Scene Parsing Network," Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 6230-6239 (2017)

[97] M. Everingham and J. Winn: "The Pascal Visual Object Classes (VOC) Challenge", International Journal of Computer Vision, Vol. 88, No.2, pp. 303-338 (2010).

[98] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang and P. A. Heng: "Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks," IEEE Journal of Biomedical & Health Informatics, Vol. 19, No. 5, pp. 1627-1636 (2015).

[99] X.- S. Wang, Y. –F. Peng, L. Le, Z. –Y. Lu, M. Bagheri and R. M. Summers: "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3462-3471, (2017).

[100] Y. –C. Wei, J. –S. Feng, X. –D. Liang, M. –M. Cheng, Z. Yao and S. –C. Yan: "Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach," Proc. of IEEE Conference on Computer Vision and Pattern Recognition, pp.

6488-6496, (2017).

[101] S. Ioffe and C. Szegedy: “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, Proc. of International Conference on International Conference on Machine Learning, Vol. 37, pp. 448-456,(2015).

[102] K. –M. He, G. Gkioxari, P. Dollár, Piotr and R. Girshick: “Mask R-CNN”, IEEE Trans on Pattern Analysis & Machine Intelligence, Vol. 99, pp. 1-1 (2017).

[103] I. J. Goodfellow, J. Pougetabadie, M. Mirza, X. Bing, D. Wardefarley, S. Ozair, Sherjil, A. Courville and Y. Bengio: “Generative Adversarial Networks”, Advances in Neural Information Processing Systems, Vol. 3, pp. 2672-2680 (2014).

[104] Y. –J. Zhu, T. Park, P. Isola and A. A. Efros: “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”, Proc. of IEEE International Conference on Computer Vision, pp. 2242-2251, (2017).

[105] J. W. Ha, Y. Cho, M. Choi, M. Kim, S. Kim and J. Choo: “StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation”, arXiv:1711.09020, (2017).

Publication List

Category	Name of publications	Reference
Journal Paper	○ Yan LI, Rong XU, Artus KROHN-GRIMBERGHE, Jun OHYA, Hiroyasu IWATA, “Deep Learning Based Uterus Localization and Anatomical Structure Segmentation on Fetal Ultrasound Image”, IIEEJ Transactions on Image Electronics and Visual Computing, (June 2019) (accepted)	Chapter 3 and Chapter 4
Journal Paper (Short Paper)	○ Yan LI, Rong XU, Artus KROHN-GRIMBERGHE, Jun OHYA, and Hiroyasu IWATA, “Region Mining of Fetal Head in Ultrasound Image Based on Weakly Supervised Annotations and Deep Learning”, IIEEJ Short Paper on Image Electronics and Visual Computing, (June 2019) (accepted)	Chapter 5
International Conferences (with reviews)	○ Yan LI, Rong Xu, Jun OHYA and Hiroyasu IWATA, “Automatic Fetal Body and Amniotic Fluid Segmentation from Fetal Ultrasound Images by Encoder-Decoder Network with Inner Layers”, EMBC, IEEE Conference on. pp, 1485-1488, (September, 2017)	Chapter 4
	○ Yan LI, Rong XU, Jun OHYA and Hiroyasu IWATA “Pregnant Uterine Ultrasound Image Segmentation by Encoding-decoding Convolutional neural network”, The 5th IIEEJ International Conference on Image Electronics and Visual Computing (March, 2017)	Chapter 4
Domestic Conference (without review)	Yan LI, Ye LI and Jun OHYA, "Road Vanishing Point Detection by Multi-Stage Convolutional Neural Network", Visual/Media Computing Conference, pp. 1-4, Tokyo, Japan, (June, 2016)	