Tissues Image Recognition in Endoscopic Surgery Using Convolutional Neural Network

CUI Peng

A DISSERTATION

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN COMPUTER SCIENCE AND ENGINEERING

Graduate Department of Computer and Information Systems

The University of Aizu

2022

The thesis titled

Tissues Image Recognition in Endoscopic Surgery Using Convolutional Neural Network

by

CUI Peng

is reviewed and approved by:

Main referee

Professor

CHEN Wenxi Wenxi Chen 陳文喜 2022/2/14

Professor

SHIN Jungpil Jungpil Shin 慎 Feb.14.2022

Professor

MARKOV Konstantin K. Markov マルコフ 2022/02/14

Senior Associate Professor

ZHU Xin Xin Zhu 朱 2022/2/15

THE UNIVERSITY OF AIZU

2022

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
AUC	Area Under Cruve
CADS	Computer-aided Detection System
CNNs	Convolutional neural networks
DL	Deep learning
DSC	Depthwise Separable Convolution
IH	Inguinal Hernia
LDH	Lumbar Disc Herniation
ML	Machine learning
PTED	Percutaneous Transforaminal Endoscopic Discectomy
ROC	Receiver Operating Characteristic
S-A	Self-Attention
SVM	Support Vector Machines
TAPP	Transabdominal Preperitoneal Patch Plasty
TEP	Total Extra-peritoneal Preperitoneal Patch Plasty

Acknowledgment

Three years of studying abroad is a long and short life journey.

I would like to dedicate my dissertation to all those who have offered me tremendous assistance during the three years in The University of AIZU.

First and foremost I am extremely grateful to my supervisor, Prof. Wenxi Chen for his invaluable advice, continuous support, and patience during my Ph.D. study. His passionate attitude towards research always inspired me. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. In my eyes, Professor Chen is a knowledgeable academic master, full of scholar temperament. I learned a lot from him, not only how to do research, but how to treat research with a scientific attitude. The advice he gave me was a valuable asset to me. I will always remember it as a navigation light for my future work and life.

Secondly, I would like to extend my sincere thanks to the other members of the committee for their constructive comments and suggestions, Prof. Jungpil Shin, Prof. Konstantin Markov, and Prof. Xin Zhu. Under the professional and patient guidance of the referees, I carefully and repeatedly revised my dissertation. Thank you very much for your comments, which have greatly improved the quality of this dissertation.

During my doctoral study, I received extensive help and support from my classmates and friends. I would like to sincerely thank them for their help and support in my study and life. I would also like to thank Mr. Wang Xiaojun and the ONO family in particular. It is their help and support in my life that makes me feel the warmth of family in a foreign country.

Finally, I would like to express my gratitude to my parents, my wife and my daughter. Without their huge support and encouragement, it would be impossible for me to complete my studies.

All in all, along the way, there are too many people worthy of my gratitude. I cannot write down their names one by one here. I will use my whole life to thank them. Thank you everyone for their encouragement and support.

Best wishes to all of you!

Abstract

Machine learning technology has broken through the bottleneck, and has made rapid progress. It is closely related to human life, especially with the wide application of artificial intelligence（AI）technology, human life is more and more convenient. With the maturity of medical technology and the continuous updating of various medical equipment, minimally invasive endoscopic surgery has been widely carried out. At the same time, the maturity of computer technology provides strong support for the medical field.

Lumbar disc herniation (LDH) is a common degenerative disease in human. In recent years, percutaneous transforaminal endoscopic discectomy (PTED) has become a major surgical treatment for LDH. In this operation, nerve and dura mater injury is one of the serious complications. One of the main reasons is that the operator cannot correctly identify the nerve and dura mater in time.

Inguinal hernia (IH) also is a common disease in human, transabdominal preperitoneal patch plasty (TAPP) and total extra-peritoneal preperitoneal patch plasty (TEP) are the main minimal invasive surgery for inguinal hernias repair, and vas deferens injury is a common complication in male patients during these minimal invasive endoscopic surgeries.

In this study, we extract the image of target tissues from endoscopic surgery videos, and use convolutional neural network (CNN) model to learn, train and test the image features of target tissues. We propose a CNN based computer-aided detection system (CADS) to complete these tasks and evaluate the performance of the CADS by various indicators.

Firstly, PTED videos of 65 patients with LDH were collected for this study. These videos were converted into images, and then the spinal endoscopy expert group labeled the areas containing nerve and dura mater in all images. The training dataset was composed of 10,454 images including nerve and dura mater of 50 patients, and 12,000 images of 15 patients constituted the test dataset. The test results showed that sensitivity (Sen), specificity (Spe), and accuracy (Acc) are 90.90%, 93.68%, and 92.29%, respectively. Compared with surgeons of different levels, the performance of CADS was higher than that of general surgeons. CADS is an effective method for nerve and dura mater recognition in PTED. With the assistance of CADS, the performance of general surgeons is close to that of spinal endoscopy experts.

Secondly, PTED videos of 102 patients with LDH were collected for this study. Eight common CNN models were used to test the recognition ability of these CNN models for nerve and dura mater images in PTED. According to the characteristics of each CNN model, the same dataset was used to train and test each model， and then the performance of these models was compared and evaluated by Sen, Spe, Acc, positive predictive value (PPV), negative predictive value (NPV), AUC and FPS. Yolov4 had the best performance in Spe 96.14% and Acc 94.42% (confidence level 0.1). The overall performance of SSD was the best, with AUC of 0.972, Sen of 92.43%, Spe of 94.21%, PPV of 94.11%, NPV of 92.56% and Acc of 93.32% (confidence level 0.3). In terms of detection speed, Yolov4-tiny was the fastest model with the detection speed 196.15 fps. CNN models can be considered as a promising method for nerve and dura mater detection under the PTED, and

Yolov4 has the best performance on this task.

Thirdly, TAPP or TEP videos of 35 patients were collected for this study, and Yolov4 was used to train and test this data. All images containing vas deferens were labeled one by one by surgery experts. The data of 26 patients were used as training data, 6 patients as image test dataset and 3 patients as video test dataset. Under the confidence level 0.4, the performance of the model in the image dataset was the best, which were TPR 90.61%, TNR 98.67%, PPV 98.57%, ACC 94.61% and F1 94.42%, respectively; In the video test dataset, the average values of TPR and TNR were 90.11% and 95.67% respectively. In this study, the vas deferens under TAPP and TEP were detected by video and image. The appropriate IoU threshold is 0.3. Yolov4 is an effective detector for vas deferens identification in TAPP and TEP.

Fourthly, in order to lighten the CNN model and apply it to our task, depthwise separable convolution (DSC) and self-attention (S-A) mechanism were used to reconstruct the Yolov4. In this study, DSC was used to replace the conventional convolution and added the S-A mechanism to the Yolov4 backbone, the parameters of the model were reduced from 63.94 $\times 10^6$ to 18.68 $\times 10^6$, the detection speed was improved from 28.12 fps to 48.84 fps. At the same time, the sensitivity was 96.21%, AUC was 0.96, which higher than Yolov4.

CNNs can accurately and effectively identify the images of nerve, dura mater and vas deferens in endoscopic surgery. DSC and S-A mechanism also can be used to reconstruct the CNN model to make the model more efficient, so that CADS can bring more convenience to clinicians, especially young doctors, and reduce unnecessary pain of patients.

概要

機械学習テクノロジーはボトルネックを突破し、急速に進歩しました。それは人間の生活と密接に関係しており、特に人工知能（AI）技術の幅広い応用により、人間の生活はますます便利になっています。医療技術の成熟と様々な医療機器の継続的な更新により、低侵襲内視鏡手術が広く行われてきました。同時に、コンピュータ技術の開発は医療分野を強力にサポートします。

腰椎間板ヘルニア(LDH)はヒトによく見られる退行性疾患です。近年、経皮経椎間孔内視鏡的椎間板切除術（PTED）がLDHの主要な外科的治療となっています。この手術では、神経と硬膜の損傷が深刻な合併症の1つです。鼠径ヘルニア（IH）も人間によく見られる疾患であり、経腹的腹膜前パッチ形成術（TAPP）と全腹膜外腹膜前パッチ形成術（TEP）は、鼠径ヘルニア修復のための主に低侵襲手術であり、輸精管損傷は男性患者がこれらの微傷内視鏡手術でよく見られる合併症である。これらの合併症の主な理由は、内視鏡下でオペレーターがこれらの組織をタイムリーかつ正確に識別できないことです。

この研究では、内視鏡手術のビデオから標的組織の画像を抽出し、畳み込みニューラルネットワーク（CNN）モデルを使用して標的組織の画像の特徴を学習します。これらのタスクを完了するためにCNNベースのコンピューター支援検出システム（CADS）を提案し、さまざまな指標によってCADSのパフォーマンスを評価しました。

最初に、LDHの65人の患者のPTEDビデオがこの研究のために集められ

ました。これらのビデオは画像に変換され、脊椎内視鏡専門家グループがすべての画像の神経と硬膜を含む領域にラベルを付けました。トレーニングデータセットは、50 人の患者の神経と硬膜を含む 10,454 枚の画像で構成され、15 人の患者の 12,000 枚の画像がテストデータセットを構成しました。テスト結果は、感度（Sen）、特異度（Spe）、および精度（Acc）がそれぞれ 90.90%、93.68%、および 92.29%であることを示しました。異なるレベルの外科医と比較して、CADS のパフォーマンスは一般外科医のパフォーマンスよりも高かった。CADS は、PTED における神経および硬膜の認識に効果的な方法です。CADS の支援により、一般外科医のパフォーマンスは脊椎内視鏡検査の専門家のパフォーマンスに近くなります。

第二に、LDH の 102 人の患者の PTED ビデオがこの研究のために集められました。8 つの一般的な CNN モデルを使用して、PTED の神経および硬膜画像に対するこれらの CNN モデルの認識能力をテストしました。各 CNN モデルの特性に応じて、同じデータセットを使用して各モデルのトレーニングとテストを行い、Sen、Spe、Acc、陽性予測値（PPV）、陰性予測値（NPV）、AUC および FPS。Yolov4 は、Acc 94.42%（信頼水準 0.1）で最高のパフォーマンスを示しました。SSD の全体的なパフォーマンスは最高で、AUC は 0.972、Sen は 92.43%、Spe は 94.21%、PPV は 94.11%、NPV は 92.56%、Acc は 93.32%でした（信頼水準 0.3）。検出速度に関しては、Yolov4-tiny は検出速度 196.15fps の最速モデルでした。CNN モデルは、PTED の下での神経および硬膜の検出のための有望な方法と見なすことができ、Yolov4 はこのタスクで最高のパフォーマンスを発揮します。

第三に、35 人の患者の TAPP または TEP ビデオがこの研究のために収集され、Yolov4 がこのデータのトレーニングとテストに使用されました。外科

専門家によって輸精管を含むすべての画像は 1 つずつマークされます。26 人の患者のデータをトレーニングデータとして使用し、6 人の患者を画像テストデータセットとして使用し、3 人の患者をビデオテストデータセットとして使用しました。信頼水準 0.4 では、画像データセット内のモデルのパフォーマンスが最高で、それぞれ TPR 90.61%、TNR 98.67%、PPV 98.57%、ACC 94.61%、F1 94.42%でした。ビデオテストデータセットでは、TPR と TNR の平均値はそれぞれ 90.11%と 95.67%でした。この研究では、TAPP および TEP 下の精管がビデオと画像によって検出されました。適切な IoU 閾値は 0.3 です。Yolov4 は、TAPP および TEP での輸精管識別のための効果的な検出器です。

第四に、CNN モデルを軽量化し、それをタスクに適用するために、深さ分離可能畳み込み（DSC）と自己注意機構（S-A）を使用して Yolov4 を再建しました。この研究では、DSC を使用して従来の畳み込みを置き換え、自己注意機構を Yolov4 バックボーンに追加し、モデルのパラメーターを 63.94×10^6 から 18.68×10^6 に減らし、検出速度を 28.12 fps から 48.84 fps に改善しました。同時に、感度は 96.21%、AUC は 0.96 で、Yolov4 よりも高かった。

CNN は、内視鏡手術で神経、硬膜、輸精管の画像を正確かつ効果的に識別できます。また、DSC および S-A メカニズムを使用して CNN モデルを再建し、モデルをより効率的にすることができます。これにより、CADS は臨床医、特に若い医師の利便性を高め、患者の不必要な痛みを軽減できます。

Chapter 1

Introduction

With the rapid development of computer technology, it has been wildly used in various field, people's daily life also more and more closely linked with the computer.

In the field of medicine, with the development of computer technology and endoscopic technology, various minimally invasive examination and minimally invasive surgery technologies have been widely used, and computer technology is more and more widely used in medical image recognition and object detection.

At present, computer technology has been widely used in clinical examination, diagnosis and treatment. It not only brings great convenience to clinicians, but also brings great benefits to patients.

1.1 Spinal Disc Herniation

1.1.1 Epidemiology

Nearly 80% of the population sustains an episode of low back pain once during their lifetime [1]. About 9%-12% of the world's population is suffering from low back pain. Low back pain is common between 40 and 80 years old. There is no significant difference in the proportion of male and female patients, and the proportion of patients will gradually increase with age [2].

Since the middle of the 20th century, the theory of low back pain caused by intervertebral disc disease has gradually become the mainstream; In the 1980s, due to the application of new technologies such as computed tomography (CT) and magnetic resonance imaging (MRI), the theory of low back pain caused by intervertebral disc disease has gained more support [3]. The illustration of spinal degeneration and disc herniation is shown in figure 1.1.

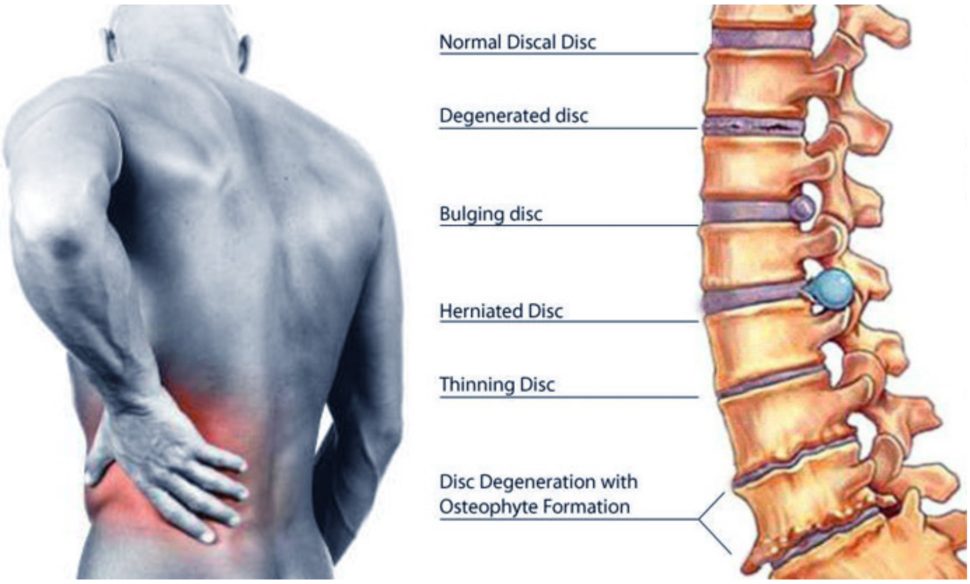


Figure 1.1: Illustration of spinal degeneration and disc herniation.

Spinal disc herniation is an injury to the cushioning and connective tissue between vertebrae, usually caused by excessive strain or trauma to the spine. It can occur in any disc in the spine, but the two most common forms are lumbar disc herniation and cervical disc herniation. The majority of spinal disc herniation occur in the lumbar spine (95% at L4-L5 or L5-S1), it occurs 15 times more often than cervical disc herniation, It may result in back pain, pain, or sensation in different parts of the body, and physical disability [4, 5].

1.1.2 Main Treatment Methods

There are many clinical treatments to alleviate the various discomfort symptoms caused by the herniation of the spinal disc. Usually, non-surgical treatment is firstly tried to treat disc herniation, but the effect of drug treatment is often not obvious, and it cannot completely solve the fundamental problem of intervertebral disc herniation, the final solution to disc herniation is surgery. In the early 20th century, American neurosurgeon Harvey Williams Cushing advocated that surgical treatment of low back pain was more accepted by the world [6].

Traditional surgical treatment has large damage to local soft tissue and bone, slow recovery, and high hospitalization cost. The choice of treatment for spinal disc herniation has always been a difficult problem for clinicians and patients.

Fortunately, with the gradual development of minimally invasive surgery technology, minimally invasive approaches to spine surgery have been pioneered and increasingly utilized over the last 15 years. These approaches are associated with small incision, less pain, less bleeding, less soft tissue and bone trauma, low nursing cost, fast recovery, decreased hospital length of stay, more and more operations including discectomy have also begun to use minimally invasive and endoscopic surgery [7–11].

The 12th (2012-2013) and 13th (2014-2015) National Surveys by the Japan Society for Endoscopic Surgery showed that endoscopic technology has been widely used in various surgery. In 2014 and 2015, 4262 patients in Japan received spinal endoscopic surgery, 3762 patients were lumbar disc herniation, accounting for 88.3% [12, 13].

Percutaneous transforaminal endoscopic discectomy (PTED) is a safe and effective method to treat lumbar disc herniation with few complications and high patient satisfaction [14, 15]. It is a minimally invasive technique to treat lumbar disk herniation from a lateral approach. Performed under local anesthesia, the incision size for PTED is around 8mm with no paraspinal muscle cutting or detachment from their insertion. In Figure 1.2, we show the real scene of PTED. We can see that experts only need to watch the computer screen to operate the equipment.

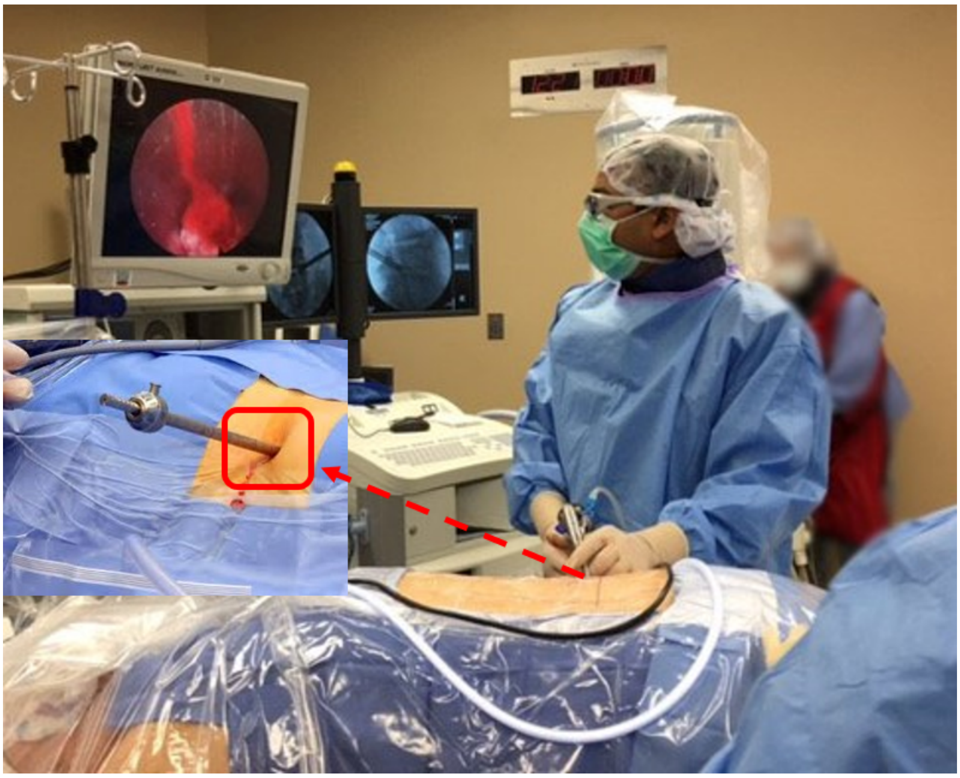


Figure 1.2: Legend of PTED.

1.2 Inguinal Hernia

1.2.1 Epidemiology

A hernia is an abnormal exit of a tissue or organ, such as the intestine, through the wall of the cavity in which it is usually located. Various types of hernias can occur, the most common being the abdomen, especially the inguina. Inguinal hernia is the most common inguinal type, but it may also be femoral type. Other types of hernia include hiatal hernia, incisional hernia and umbilical hernia. About 66% of patients with inguinal hernia have symptoms. This may include lower abdominal pain or discomfort, especially in or when cough, exercise, urination or defecation [16].

Inguinal hernia is a protrusion of abdominal-cavity contents through the inguinal canal. For men, there is a lifelong risk of contracting an inguinal or femoral hernia of 27%-43%. For women, there is a risk of 3%-6% [16,17]. In young infants, the incidence of indirect inguinal hernia is approximately 1% to 5%, with male to female ratio of 10:1 [18].

In 2015, inguinal, femoral and abdominal hernias affected about 18.5 million people [19]. In the western countries, including the United States, more than 1.5 million procedures are performed every year [20]. Globally, inguinal, femoral and abdominal hernias resulted in 60,000 deaths in 2015 and 55,000 in 1990 [21, 22]. Inguinal hernia repair is one of the most frequent operation in general and visceral surgery worldwide.

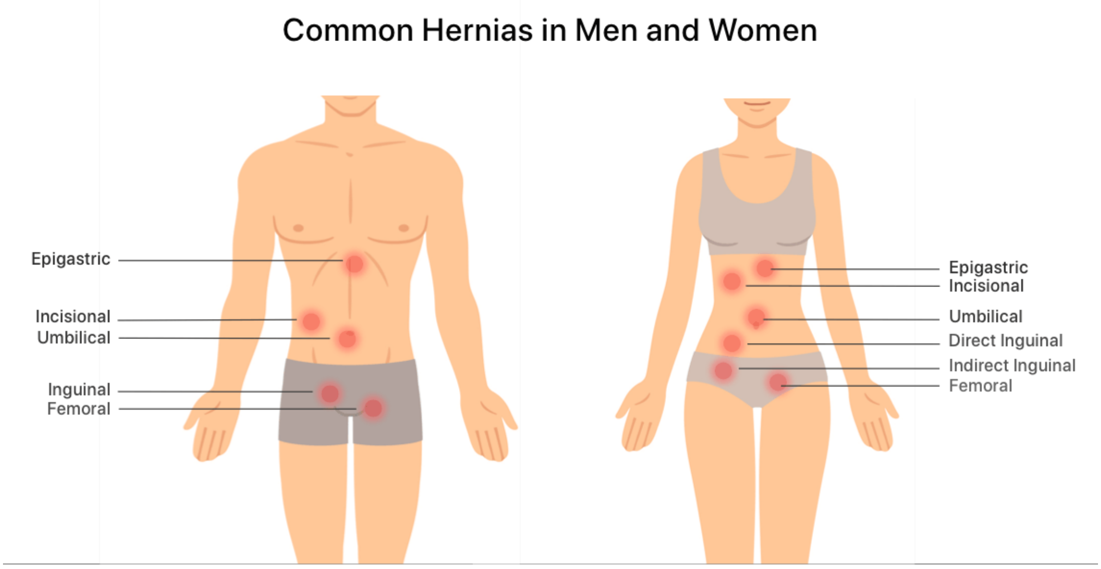


Figure 1.3: The common hernias in human.

(Picture source: https://www.drugwatch.com/wp-content/uploads/Common-Hernias–640x0-c-default.webp)

1.2.2 Main Treatment Methods

Bassini performed the first inguinal hernia repair with reconstruction of the floor of the inguinal canal to preserve functional anatomy in 1884, firstly described in 1887 [23].

In 1982, under the guidance of laparoscopic technology, Ger et al. [24] used laparoscopic technology to create a new type of hernia repair. Two types of laparoscopic surgery are successfully applied: the transabdominal preperitoneal patch plasty (TAPP), and the totally extra-peritoneal preperitoneal patch plasty (TEP).

TAPP and TEP are technically difficult and require special skills to overcome the inherent limitations of such surgery. It has a significant learning curve and is related to prolonging the operation time [25, 26]. But in the hands of trained surgeons, it is safe and effective.

Compared with the traditional open surgery, TAPP and TEP may provide better posterior inguinal wall exposure, enabling easier bilateral reinforcement, it also can reduce the length of hospital stay and earlier return to work [27, 28].

Figure 1.4: Legend of laparoscopic surgery.

1.3 Development of Convolutional Neural Networks

As an important part of computer technology, artificial intelligence (AI), machine learning (ML) and deep learning (DL) have penetrated into every corner of human life in recent years.

In Figure 1.5, we briefly describe the relationship among the three technologies.

1.3.1 AI, ML, DL, and CNNs

1.3.1.1 Artificial Intelligence

At the Dartmouth conference in 1956, the concept of "AI" was first put forward. In the following ten years, artificial intelligence ushered in the first small peak

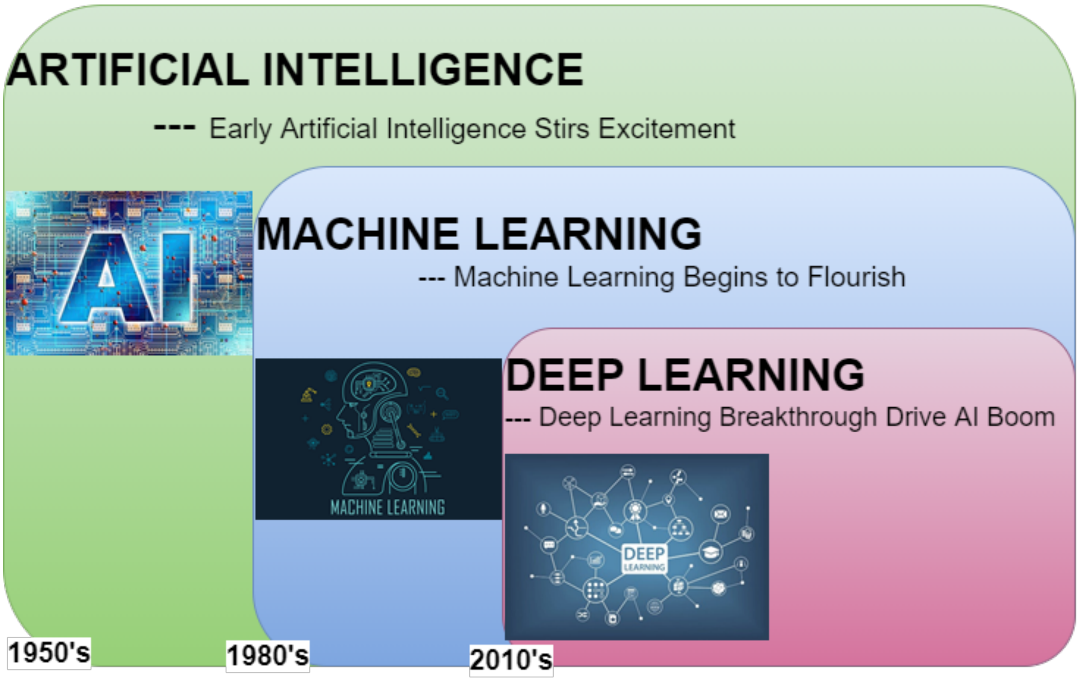

ARTIFICIAL INTELLIGENCE
--- Early Artificial Intelligence Stirs Excitement
AI
MACHINE LEARNING
--- Machine Learning Begins to Flourish
MACHINE LEARNING
DEEP LEARNING
--- Deep Learning Breakthrough Drive AI Boom
DEEP LEARNING
1950's
1980's
2010's

Figure 1.5: The origin and development of artificial intelligence

in the history of development. Researchers swarmed in and made a number of remarkable achievements.

AI means that computers have the same intelligence ability as human beings. It is a frontier comprehensive subject integrating computer science, statistics, brain neuroscience and social science. It can replace human beings to realize recognition, cognition, analysis, decision-making and other functions. [29]

1.3.1.2 Machine Learning

ML is seen as a subset of artificial intelligence. It is a multi-disciplinary interdisciplinary, involving probability theory, statistics, approximation theory and complex algorithms. It uses computer as a tool and is committed to simulating or realizing human learning behavior, and divides the existing content into knowledge structure to effectively improve learning efficiency. [30]

At present, ML technology plays an important role in medical image processing, image retrieval and analysis, computer-aided diagnosis, image interpretation, image fusion, image registration, image segmentation, image-guided therapy and other image recognition fields.

ML algorithms are often categorized as supervised learning, unsupervised learning, and reinforcement learning. ML techniques are applicable to addressing various problems, such as classification, regression, clustering, and dimension reduction. We summarized the details in Figure. 1.6.

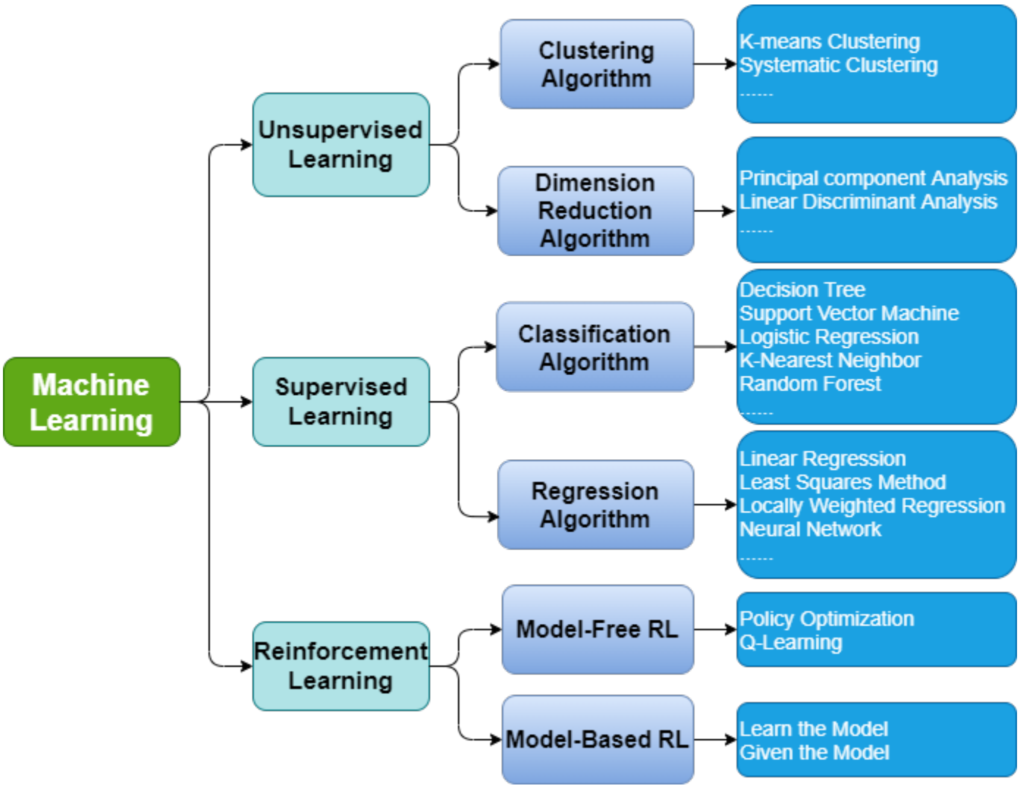


Figure 1.6: Machine learning technology

1.3.1.3 Deep Learning

DL is a branch of ML. It is a complex machine learning algorithm, which is used to learn the inherent laws and representation level of sample data. Although DL is a kind of ML, DL is to use deep neural network to make the model more complex, so as to make the model understand the data more deeply.

The information obtained in the learning process is of great help to the interpretation of text, image, sound and other data. Its ultimate goal is to enable the machine to have the same analysis and learning ability as human beings, and to recognize data such as words, images and sounds [31].

DL methods can also be divided into supervised learning, semi-supervised learning, and unsupervised learning. There are great differences in learning models established under different learning frameworks. For example, CNN is a machine learning model under deep supervised learning, while deep belief network (DBN) is a machine learning model under unsupervised learning.

DL can process a large number of data, extract features automatically, and solve complex problems. It shows good adaptability and strong learning ability in various applications. Deep learning architecture, such as deep neural network (DNN), deep belief network (DBN) and recurrent neural network (RNN), has been applied to computer vision, speech recognition, natural language processing, material detection and other fields. Their results are comparable to human experts, and even better than human experts in some cases.

Of course, DL is not perfect. It depends on data, and its interpretability is not high. DL needs a lot of data and a lot of computing power, which leads to its high requirements for hardware, so the cost is very high. At the same time, the design of deep learning model is very complex, which requires a lot of manpower, material resources and time to develop new algorithms and models.

1.3.1.4 Convolutional Neural Network

The development of CNNs can be traced back to 1962, originated from Hubel and Wiesel's research on the visual system in cat brain. In 1980, Japanese scientist Fukushima Bangyan proposed Neocognitron, a neural network structure including convolution layer and pooling layer [32]. In 1998, on this basis, Yann LeCun [33] proposed LeNet, and applied BP algorithm to the training of this neural network structure, which formed the initial prototype of convolution neural network. With the proposal of ReLU and Dropout, as well as the historical opportunities brought by GPU and big data. In 2012, CNN, represented by AlexNet [34], ushered in a historical breakthrough and entered a stage of explosive development. The development of CNNs is briefly shown in Figure 1.7.

CNNs is a kind of feed-forward neural network. Its artificial neurons can

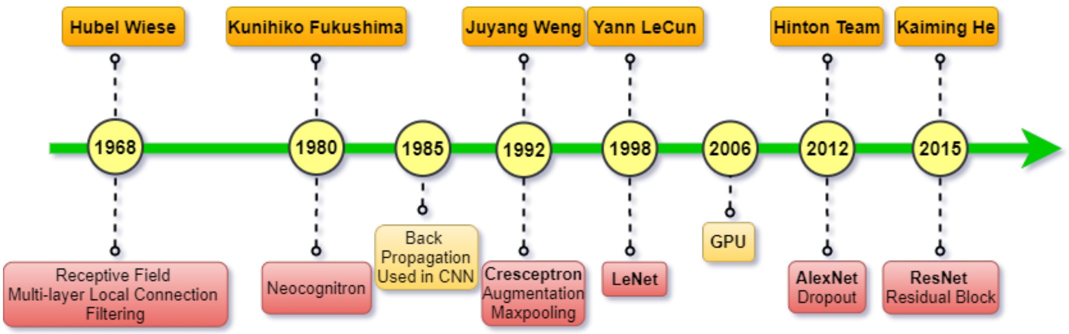


Figure 1.7: History and development of CNN

respond to some surrounding cells in the coverage area, and it has excellent performance for large-scale image processing. It is mainly composed of these layers: input layer, convolution layer, ReLU layer, pool layer and full connection layer. By superposing these layers, a complete convolutional neural network can be constructed. In Figure 1.8, we show the basic architecture of CNNs.

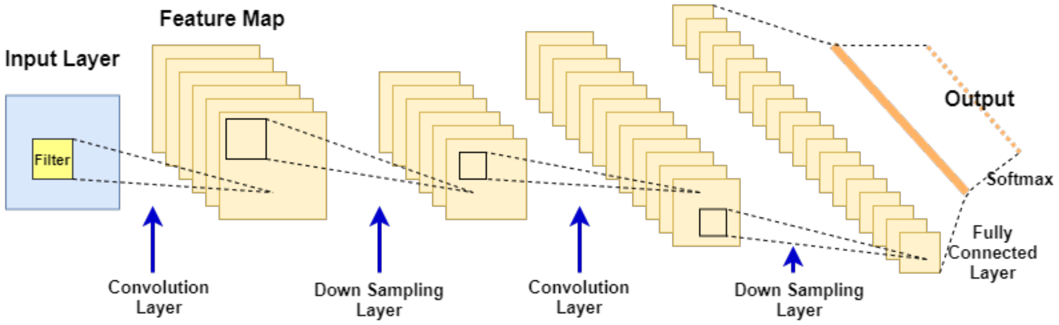


Figure 1.8: The basic architecture of CNNs

1.3.2 Common CNN Models in Object Detection

According to the difference of model structure, these object detection models are generally divided into two categories, two-stage and one-stage detector.

1.3.2.1 Two-stage Detectors

The structure of the two-stage detector is complex, which mainly focuses on the accuracy of object detection. [35–37] Two-stage detectors decouple the task of object localization and classification for each bounding box. It firstly filters out the regions that have high probability to contain an object from the entire image with region proposal network (RPN). [35] Then the proposals are fed into the region convolutional neural network (R-CNN) [38] and get their classification score and the spatial offsets. A simple two-stage algorithm is shown in Figure 1.9.

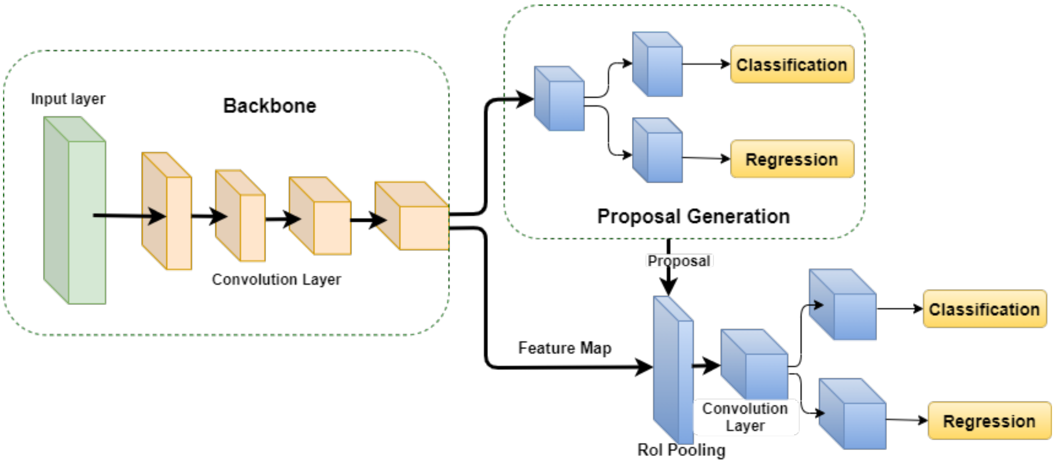


Figure 1.9: Two-stage algorithm architecture.

For now, two-stage detectors take the lead in detection accuracy. In these detectors, sparse region proposals are generated in the first stage and then are further regressed and classified in the second stage. RCNN [39] utilized low-level computer vision algorithms such as Selective Search [40] and Edge Boxes [41] to generate proposals, then adopt CNN to extract features for training SVM classifier and bounding box regressor. SPP-Net [42] and Fast R-CNN [38] proposed to extract features for each proposal on a shared feature map by spatial pyramid pooling. Faster R-CNN [35] integrated the region proposal process into the deep convnet and makes the entire detector an end-to-end trainable model.

1.3.2.2 One-stage Detectors

One-stage detector pays more attention to detection speed. One-stage detectors make the predictions for object localization and classification at the same time. These kind of detectors [43–47] usually adopt a straightforward fully convolutional architecture, and the outputs of the network are classification probabilities and box offsets at each spatial position. A simple one-stage algorithm is shown in

Figure 1.10.

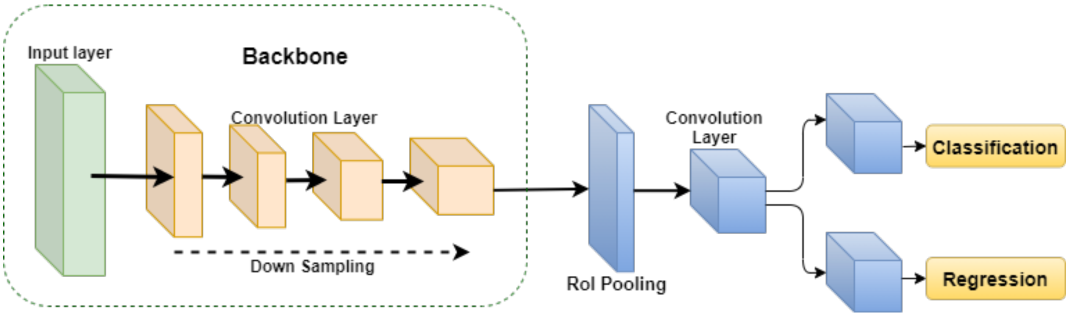


Figure 1.10: One-stage algorithm architecture.

One-stage detectors perform classification and regression on dense anchor boxes without generating a sparse RoI set. Yolo [48] is an early exploration that directly detects objects on dense feature map. SSD [49] proposed to use multi-scale features for detecting variant scale objects. RetinaNet [44] proposed focal loss to address the extreme class imbalance problem in dense object detection. RefineDet [46] anchor refinement module and the object detection module to imitate the cascade regression on dense anchor boxes. Guided Anchor [50] first used anchor-guided deformable convolutions to align features for RPN.

We have arranged these common algorithms according to the time sequence of their emergence, as shown in Figure 1.11.

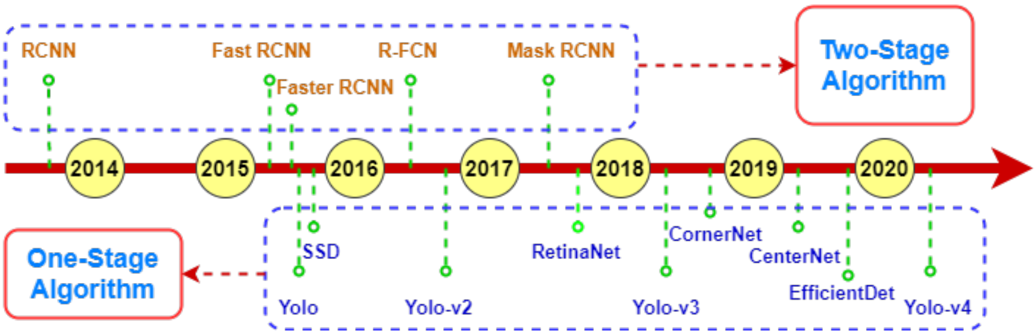


Figure 1.11: Common Two-stage and One-stage CNN models

1.4 Computer Aided Detection/Diagnosis System in Medical Field

Medical image analysis aims to provide clinicians and radiologists with a more effective diagnosis and treatment process. However, with the development of science and technology, in the traditional computer-aided detection/diagnosis (CAD) system, manual interpretation of data has gradually become a challenging task. Deep learning methods, especially CNN, have been successfully used as a tool to solve this problem.

Doctors may misunderstand the disease due to lack of experience or fatigue, resulting in missed diagnosis. Non lesions can be interpreted as lesions, or benign lesions are misunderstood as malignant ones. According to statistics, in medical image analysis, the misdiagnosis rate caused by human factors can reach 10-30% [51]. In this context, for doctors, CAD system is a very useful tool in medical image analysis.

The application of CNN-based method in medical images is very different from that in natural images [42]. In the training and testing of CNN, a large-scale labeling dataset is needed, which requires experienced experts to complete a lot of labeling works, so large-scale medical image datasets are not always available.

CNN-based methods are actively used for tasks such as classification, localization, segmentation and registration in medical image analysis. To the clinicians and radiologists, it is not the separation or combination of these tasks but the incorporation of them with a unified system that plays a significant role in clinical applications, known as the CAD systems.

CAD is a concept established by taking into account equally the roles of physicians and computers, whereas automated computer diagnosis is a concept based on computer algorithms only. With CAD, the performance by computers does not have to be comparable to or better than that by physicians, but needs to be complementary to that by physicians [52]. In fact, since 1960s, a large number of CAD systems have been used to assist doctors in the early detection of breast cancers on mammograms. [53]. In 1998, the first equipment approved by the U.S. Food and Drug Administration (FDA) was a breast X-ray CAD equipment manufactured by R2 technology. The survey shows that in 2008, 74% of screening mammograms used CAD [54], and in 2016, 92% of screening mammograms used CAD [55].

Figure 1.12 shows three common types of CAD systems [56]. a. First-reader type: After pre-selection by CAD system, the results are presented to doctors for further processing. b. Second-reader type: Doctors first read the pictures to be examined, and CAD will also give results to the pictures to be examined. In the following steps, the doctor refers to the results of CAD and determines whether each CAD result has neglected lesions or false-positive results. c. Concurrent-reader type: The doctor reads the pictures to be examined and displays the CAD results at the same time. The doctor can accept or reject the CAD results and combine them with his / her own results without reading them again.

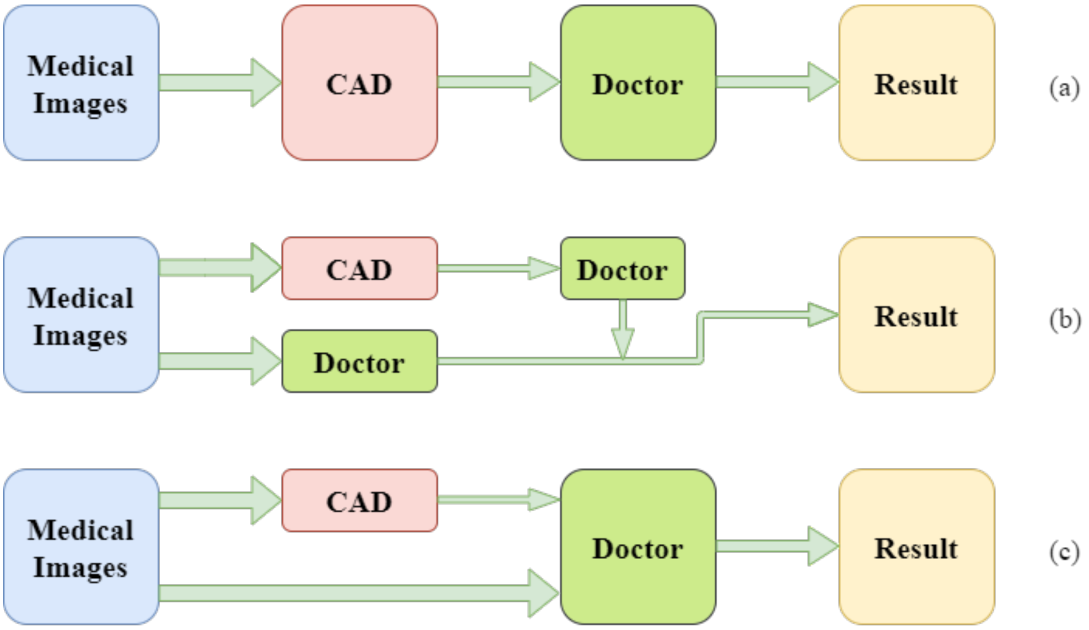


Figure 1.12: Three typical CAD system workflows:(a) First-reader type. (b) Second-reader type. (c) concurrent-reader type.

Currently, CAD systems are widely used for the detection and diagnosis of diseases in medical image analysis, such as breast cancer, lung cancer, gastrointestinal diseases, prostate cancer, osteoporosis analysis, skin lesions, Alzheimer's disease, COVID-19, and so on. The application of CAD systems can improve the accuracy of diagnosis, reduce time consumption, and optimize the doctors' workloads and learning curve.

1.5 Thesis Structure

This dissertation is a report about three years of my research on the topics of tissues recognition under various kinds of minimally invasive endoscopic surgery using CNN.

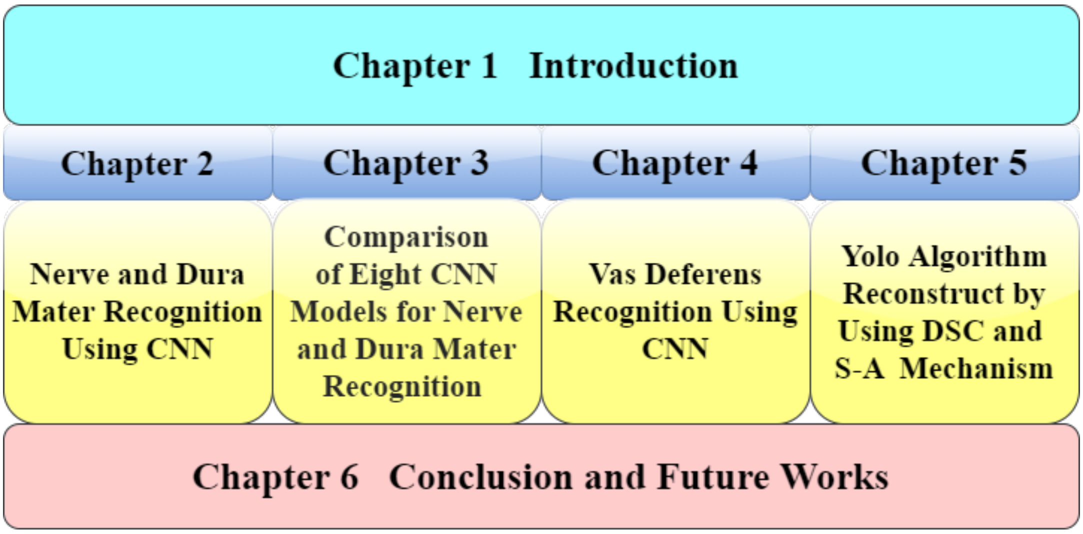

Chapter 1 Introduction
Chapter 2
Chapter 3
Chapter 4
Chapter 5
Nerve and Dura Mater Recognition Using CNN
Comparison of Eight CNN Models for Nerve and Dura Mater Recognition
Vas Deferens Recognition Using CNN
Yolo Algorithm Reconstruct by Using DSC and S-A Mechanism
Chapter 6 Conclusion and Future Works

Figure 1.13: Organization of thesis structure

This dissertation mainly consists of six parts as shown in Fig. 1.13. At first, some background knowledge is introduced in Chapter 1, including the relate knowledge of spinal vertebral disc herniation and inguinal herniation, introduced development of artificial intelligent, and computer-aided detection/diagnosis system in medical image analysis field. Additionally, the thesis structure and the main contributions are also summarized in this chapter. In Chapter 2, for the nerve and dura mater recognition, we used CNN models to train and validate 65 patients' data extracted from PTED, and estimated the performance of CNN model and different level surgeons. Further, in Chapter 3, for comparing the performance of different CNN models, we used the same training data, test data and similar parameters to train and test 8 popular CNN models. In Chapter 4, we extend the research object, we collected vas deferens images from laparoscopic hernial repair surgery, and used CNN model to recognize the image of vas deferens. In Chapter 5, We reconstruct the CNN model based on Yolov4. DSC is used to replace the conventional convolution in the Yolov4 backbone, and S-A mechanism is added in the preliminary feature extraction process to reconstruct the CNN model, and satisfactory results are obtained. Finally, in Chapter 6, the thesis is concluded and future works are presented.

1.6 Main Contributions

In this thesis, I focused on the identification of various tissues under different endoscopic operations. I used CNNs to identify the nerve and dura mater under PTED and the vas deferens under laparoscopic surgery (TAPP, TEP). I also tried to reconstruct CNN model based on Yolov4 by using depthwise separable convolution (DSC) and self-attention (S-A) mechanism. Chapters 2, 3, 4 and 5 described the methodology and results. The main contributions of each chapter are summarized as follows.

1. In Chapter 2. Nerve and dura mater recognition under PTED using CNN

   - The CNN model Yolov3 was used to train and test the nerve and dura mater datasets, and the performance of the model was evaluated. The performances of CNN model and surgeons at different levels were compared and used as another evaluation index to evaluate the performance of CNN model. The results of CNN model were combined with each surgeon, and then the effect of CNN model was evaluated by statistical method.

   - This is the first time that CNN model has been used to identify the nerve and dura mater under PTED. The experimental results shown that CNN can recognize tissue images in the process of pted, which is an effective nerve and dura mater recognition method, and can assist general surgeons in the process of PTED.

2. In Chapter 3. Comparison of 8 CNN models in detecting nerve and dura mater under PTED

   - In order to further verify whether CNN model can effectively identify nerve and dura mater in PTED, eight different CNN models (including one-stage and two-stage) was used to detect and identify nerve and dura mater in PTED.

   - Using the same index, the nerve and dura mater recognition performance of eight CNN models in PTED was evaluated and compared. The results shown that these CNN models can effectively identify nerve and dura mater. The best AUC was 0.972, produced by SSD, the best precision was 94.42%, produced by Yolov4, the fastest model was Yolov4-Tiny, the detection speed can reach 196.15 fps.

   - The function of CNNs has been proved to be an effective tool for detecting nerve and dura mater. It can assistant surgeons accurately and timely identify nerve and dura mater in PTED.

3. In Chapter 4. Vas deferens recognition under Laparoscopic using CNN

   - Yolov4 was used as a basic model to identify vas deferens images under laparoscopic hernia repair.

   - Different labeling methods were used to label the training data, which provides more effective training features for the training model, so as to improve the training effect and test accuracy of the model.

   - The real-time detection ability of the model was tested by using video clips (25 fps). The test results shown that the CNN model can recognize the vas deferens in images and videos.

4. In Chapter 5. Yolo algorithm based on self-attention mechanism and depthwise separable convolution

   - In order to lighten the model, depthwise separable convolution (DSC) was used to replace the conventional convolution in the backbone of Yolov4. At the same time, in order to increase the receptive field in the process of feature extraction, self-attention (S-A) mechanism was added to the step of preliminary feature extraction.

   - The new models were tested, and the performance of the new model was compared from the aspects of model size, detection speed, area under curve (AUC), accuracy and so on.

   - The results shown that after using DSC combined with S-A mechanism in the preliminary feature extraction process, the number of parameters of the model was reduced to one third of Yolov4, the detection speed reached 48.84 fps, and the overall detection performance remained at the level of original Yolov4.

1.7 Publications

The following papers have been published by major journals and conferences.

1. Major journal papers

   - **Peng Cui**, Tao Shu, Jun Lei, Wenxi Chen. Nerve recognition in percutaneous transforaminal endoscopic discectomy using convolutional neural network [J]. Medical Physics, 2021, 48(5): 2279-2288.

   - **Peng Cui**, Song Zhao, Wenxi Chen. Identification of vas deferens in laparoscopic inguinal hernia repair surgery using convolutional neural network [J]. Journal of Healthcare Engineering, vol. 2021, Article, 10 pages, 2021.

2. Major conference paper

- **Peng Cui**, Zhe Guo, Jianbo Xu, Tianhui Li, Yuchen Shi, Wenxi Chen, Tao Shu, Jun Lei. Tissue Recognition in Spinal Endoscopic Surgery Using Deep Learning. 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), IEEE, 2019:1-5.
- Jianbo Xu, **Peng Cui**, Wenxi Chen. ECG-based Identity Validation during Bathing in Different Water Temperature. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), IEEE, 2020: 5276-5279
- Tianhui Li, Jianbo Xu, **Peng Cui**, Wenxi Chen. The Effect of Stress on Optimal Bathing Time. 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), IEEE, 2019:1-5.
- Jianbo Xu, Tianhui Li, **Peng Cui**, Wenxi Chen. Improvement of ECG based Personal Identification Performance in Different Bathtub Water Temperature by CNN. 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), IEEE, 2019:1-5.

Chapter 2

Nerve and Dura Mater Recognition under PTED Using CNN

2.1 Introduction

Lumbar disc herniation is a common cause of low back pain and lower limb radiation pain. Conservative treatment can improve the symptoms in most cases. However, in some cases, after conservative treatment, the pain is still continuous or recurrent after alleviating for a period of time. In this case, surgical treatment will be considered [57, 58].

Among the surgical methods for LDH, open lumbar microdiscectomy （OLD） is considered the gold standard [59]. Recently, PTED is also commonly performed for lumbar disc herniation for its various strong points.

Although spinal endoscopic surgery has more advantages [60], it also has its special limitations. Due to the decrease of tissue exposure, narrow working channel and limited visual field during the operation, the operators need to have rich experience in the operation, so the operation is relatively difficult, and there will be some special complications [61, 62].

There are several important complications associated with discectomy for LDH. The rate of dura mater tears following LDH ranges from 1% to 17% and is increased particularly with ageing, obesity, and revision procedures [63]. Other complications include post-operative infection (1-5%), worsening of functional status (4%), and nerve root injury (0.2%) [64, 65]. The 2011 nationwide spinal surgery complications survey in Japan showed that nerve root damage, spinal cord damage, cauda equina damage and dura mater tear, cerebrospinal fluid leakage are the most common complications of the intraoperative and postoperative [66].

Surveys have shown that about 11.39% of patients have had various complications during endoscopic spinal surgery, and dura tear was the most common complication of endoscopic spinal surgery, accounting for about 2.7% [67]. M. J. Perez-Cruet et al. [68] analyzed the causes of the complications of spinal endoscopic surgery and the operation effects of different qualifications of spinal endoscopic doctors. The investigation shows that the operation of endoscopic surgery requires the rich experience and skills of surgeons. One of the difficulties in spinal endoscopic surgery is that surgeons need to be very familiar with the anatomy of the spine and recognize various tissues under the endoscope. Before the surgeons overcome these barriers, the incidence of complications is higher [69–72].

The innovations of this study are as follows: (1) A computer-aided detection system (CADS) is proposed to recognize the nerve and dura mater images in PTED; (2) The detection results of CADS were compared with those of doctors at different levels, and the performance of CADS was evaluated by sensitivity, specificity, accuracy and statistical methods; (3) The misrecognized and unrecognized images of CADS and clinicians are analyzed, and reasonable suggestions were provided for clinicians.

2.2 Materials and Methods

This research was approved by the Ethics Committee at Zhongnan Hospital (NO.2020066K), Wuhan University, Hubei Province, China. All patients were requested to satisfy some conditions and sign written informed consent before participation. To protect patient privacy, we only collected information on age, gender, illness, surgical location of the patient, and the video of the PTED as research data. All of the endoscopic surgeries were performed by senior spinal endoscopic experts at Zhongnan Hospital.

The inclusion criteria were as follows: (1) patients were undergoing spinal surgery for the first time (including open spinal surgery and minimally invasive surgery); (2) patients agreed to choose PTED as the surgical method; and (3) patients agreed for the surgical video to be used for scientific research. The following patients were excluded: (1) patients with a history of surgery at this site; (2) patients who had undergone spinal endoscopic surgery due to intervertebral infection, spinal cord tumor, radiculopathy, and other diseases; and (3) patients who did not agree for the surgical video to be used for scientific research.

2.2.1 Datasets

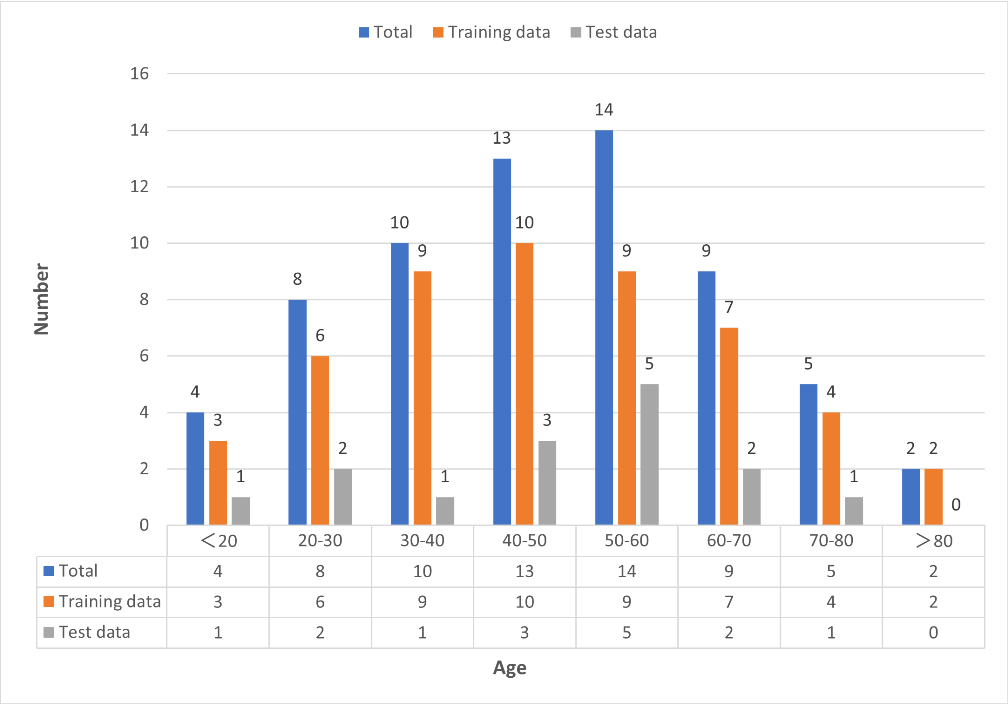
In total, we collected videos from 65 patients (46 males/19 females; age range, 17-87 years, mean 48.06 ± 16.91) with LDH who underwent PTED at Zhongnan Hospital from January to December 2019. All patients had different levels of lumbocrural pain and limb numbness. Detailed information is given in Table 2.1 and the patient age distribution is shown in Figure 2.1.

2.2.1.1 Training Dataset and Validation Dataset

Firstly, we randomly selected videos from 50 patients as the training sample; we selected video clips with nerve and dura mater from these videos, and then used MATLAB (9.6.0.1174912 (R2019a) Update 5, academic use) to convert video into

Table 2.1: Patient information

Dataset	Age (years)	Gender		Site of lesion		
	Mean ± SD	Male	Female	L3/4	L4/5	L5/S1
Training	48.34 ± 17.56	36	14	09	23	18
Test	47.13 ± 15.06	10	05	02	08	05



	<20	20-30	30-40	40-50	50-60	60-70	70-80	>80
■ Total	4	8	10	13	14	9	5	2
■ Training data	3	6	9	10	9	7	4	2
■ Test data	1	2	1	3	5	2	1	0

Figure 2.1: Patient age distribution.

images. Three spinal endoscopic experts using LabelImage manually checked and labeled all the nerve and dura mater images one by one. In total, 10,000 images (200 per patient) were selected and labeled.

To provide more features for CNN learning, we also extracted 454 images that were not correctly identified by the CNN model in the "test data" in our previous experiments. [73] The same three experts manually labeled the position of nerve and dura mater in these images. They then examined all the images and labels in the dataset, discussed some of the details, and reached a consensus on the labeling of each image.

We added these 10,454 images as input data into the new training dataset and randomly divide these images into training data and validation data in a ratio of 8:2. We adjusted the model parameters, and then used these data as input data to start training and to verify the model. In this process, we used the weight of pretraining to improve the quality and speed of training. When the verification accuracy reached the optimal, the model parameters were automatically saved. Sample images of the original data and the the labeled data are shown in Figure 2.2.

2.2.1.2 Test Dataset

For the test dataset, we selected 12,000 images from the 15 patients (400 images containing nerve and dura mater and 400 images without nerve and dura mater taken from each patient). Neither the patients nor the images in this dataset overlapped with the training and validation datasets. All images in the test dataset were verified by three spinal endoscopic experts one by one.

2.2.1.3 Computer Configuration

All the original data come from surgical videos of patients undergoing PTED. The endoscope used was from Richard Wolf GmbH, 75438 Knittlingen, Germany. The computer comprised an Intel i5 3570k 3.4-GHz CPU, with 4TB hard disk

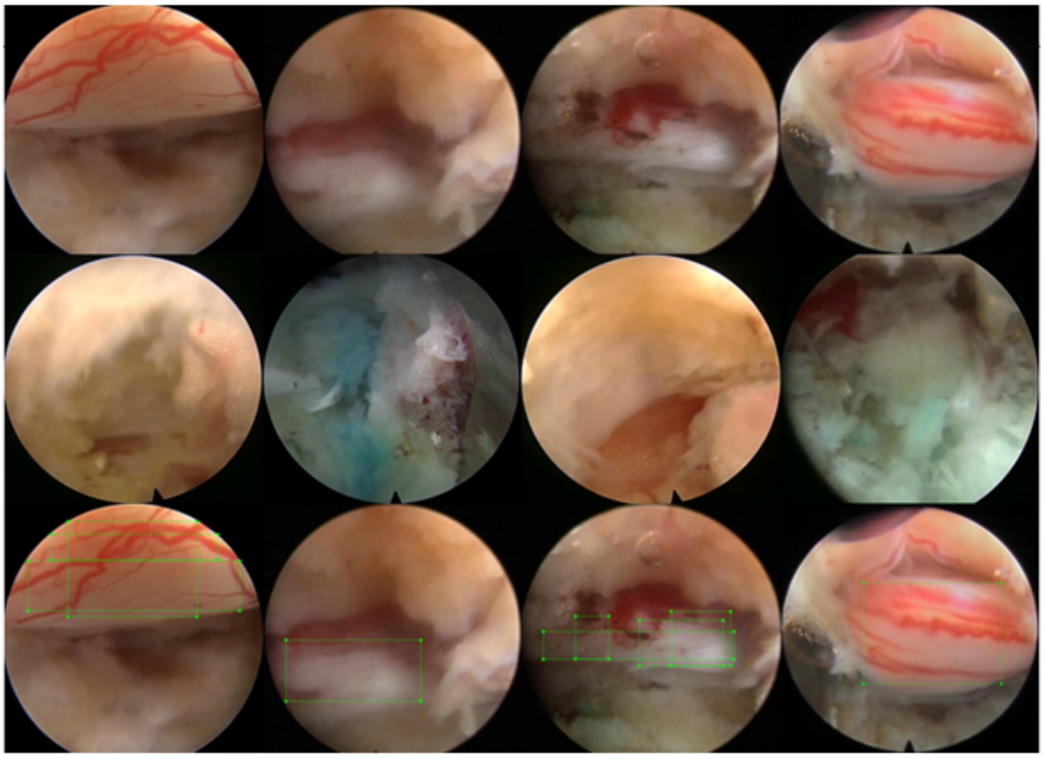


Figure 2.2: The first row shows the original image containing nerve and dura mater under the PTED view, with clear, unclear, pale, and ruddy images; the second row shows the original image without nerve and dura mater, including clear, unclear, and easily confused with nerve and dura mater; the third row shows images of nerve and dura mater, labeled with green rectangles by the spinal endoscopic experts.

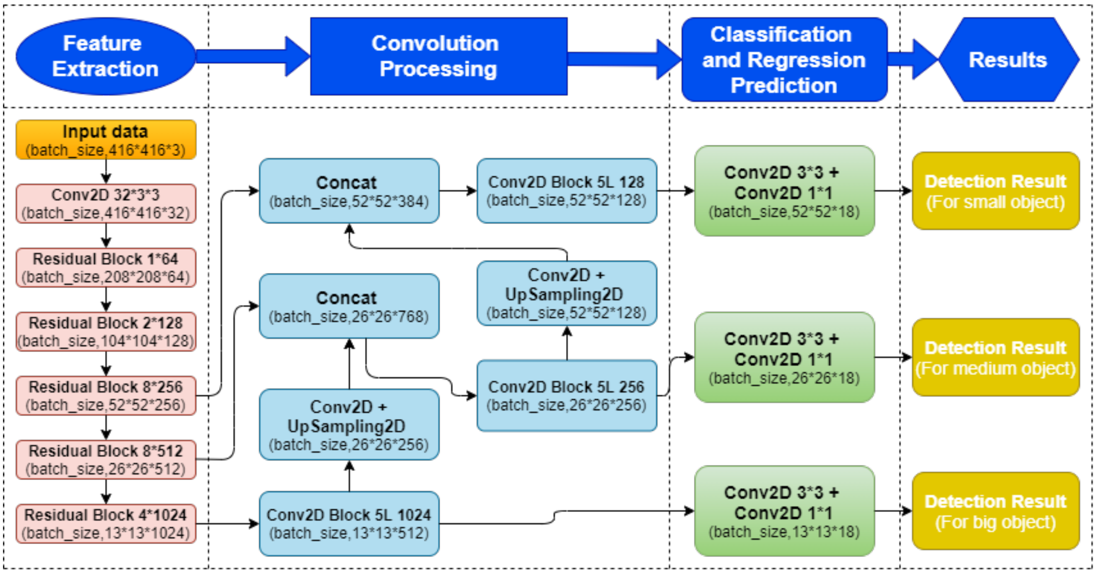

Feature Extraction
Convolution Processing
Classification and Regression Prediction
Results
Input data (batch_size,416*416*3)
Conv2D 32*3*3 (batch_size,416*416*32)
Residual Block 1*64 (batch_size,208*208*64)
Residual Block 2*128 (batch_size,104*104*128)
Residual Block 8*256 (batch_size,52*52*256)
Residual Block 8*512 (batch_size,26*26*512)
Residual Block 4*1024 (batch_size,13*13*1024)
Concat (batch_size,52*52*384)
Conv2D Block 5L 128 (batch_size,52*52*128)
Concat (batch_size,26*26*768)
Conv2D + UpSampling2D (batch_size,52*52*128)
Conv2D + UpSampling2D (batch_size,26*26*256)
Conv2D Block 5L 256 (batch_size,26*26*256)
Conv2D Block 5L 1024 (batch_size,13*13*512)
Conv2D 3*3 + Conv2D 1*1 (batch_size,52*52*18)
Conv2D 3*3 + Conv2D 1*1 (batch_size,26*26*18)
Conv2D 3*3 + Conv2D 1*1 (batch_size,13*13*18)
Detection Result (For small object)
Detection Result (For medium object)
Detection Result (For big object)

Figure 2.3: CNN Model Structure.

space, 32 GB RAM, and a CUDA-enabled Nvidia Titan 312GB graphics processing unit (Nvidia), based on the hardware of the Nvidia GeForce 2080Ti Pascal GPU.

2.2.2 CNN Model and CADS Work Flow

The CNN model we used is a one-stage algorithm, which applies a single neural network to the whole image, and uses CNN to directly predict the location of different targets and classify them. The algorithm treated the problem of target detection as a regression problem. Using the structure of CNN, the bounding box and class probability can be predicted directly from the input image. Through the combination of reinforcement learning and supervisory learning, the stability and sustainability of CADS are guaranteed. A simplified model structure and model diagram of CADS are shown in Figure 2.3 and Figure 2.4.

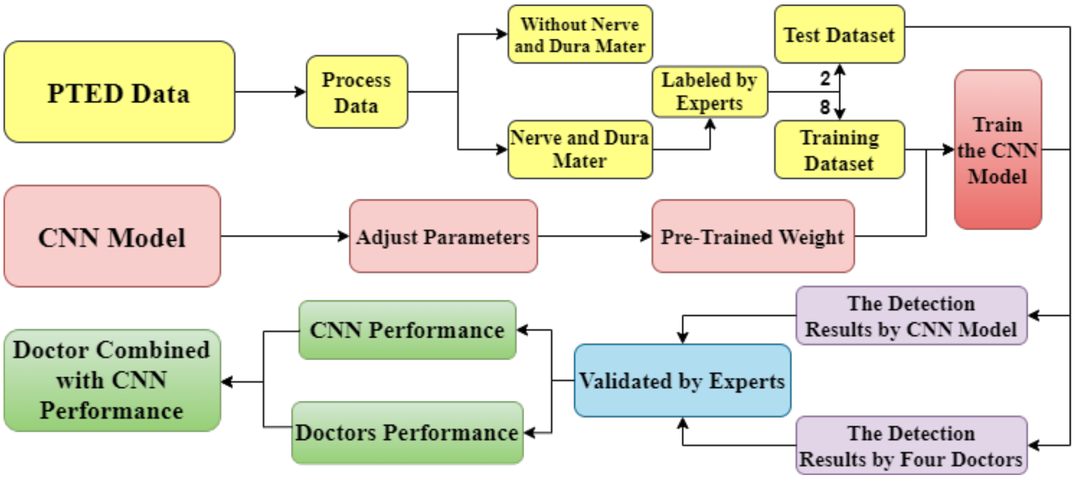


Figure 2.4: Flowchart for CADS.

2.3 Results

All images in the test dataset were identified and labeled by the trained model, and then verified by three senior spinal endoscopic experts one by one.

2.3.1 Test Dataset

This model correctly identified 11,075/12,000 (92.29%) images of 15 patients, including 5,454/6,000 (90.90%) images with nerve and dura mater, and 5,621/6,000 (93.68%) images without nerve and dura mater. To analyze the results more objectively, we also analyzed the sensitivity, specificity, and accuracy of the model in test dataset for nerve and dura mater recognition of each patient. The images are shown in Figure 2.5, and the results are shown in Table 2.2.

2.3.2 Training and Testing Process

We adjusted the parameters in the training process and set the image input size to 416*416, the initial learning rate = 0.001, the number of iterations = 20000. We add the verification accuracy and loss curves in the training process, and the receiver operating characteristic (ROC) curve in the test process to observe the performance of the model. When the confidence threshold is 0.3, the average intersection over union (IoU) of the validation dataset is 81.41%, and that of the

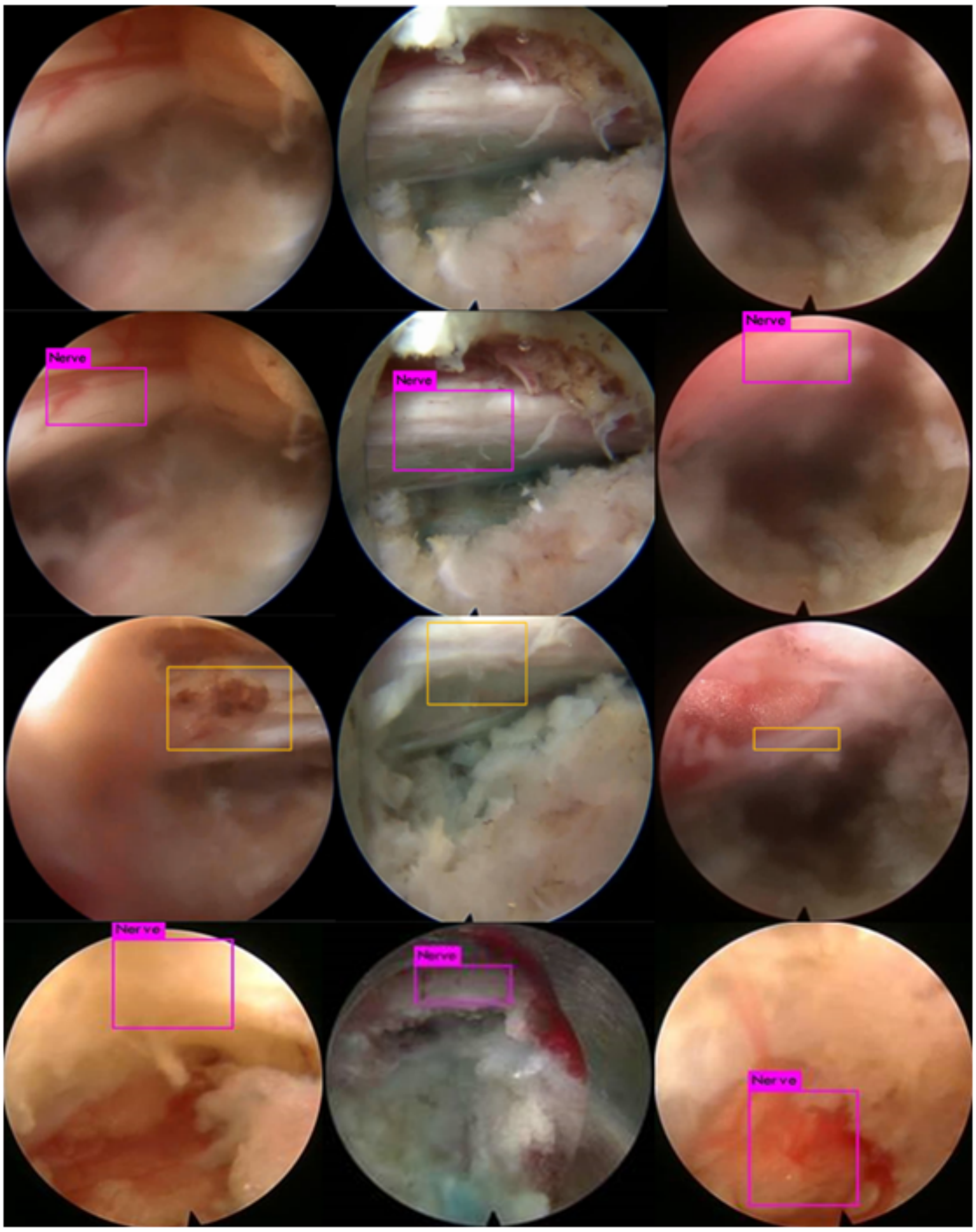


Figure 2.5: The top row shows original images in the test dataset. The second row shows the true positives (TP), which labeled the nerve or dura mater (purple rectangles) correctly. The third row shows the false negatives (FN), with nerve or dura mater in the image, which the model misidentified and did not label. We labeled nerve or dura mater with yellow rectangles. The bottom row shows the false positives (FP), with no nerve or dura mater in the image, which the model misrecognized and mislabeled (purple rectangles).

Table 2.2: Individual performance of test dataset

Patient	TP	TN	FP	FN	Sen (%)	Spe (%)	Acc (%)
1	368	374	26	32	92.00	93.50	92.75
2	361	378	22	39	90.25	94.50	92.38
3	372	377	23	28	93.00	94.25	93.63
4	359	375	25	41	89.75	93.75	91.75
5	366	372	28	34	91.50	93.00	92.25
6	351	388	12	49	87.75	97.00	92.38
7	365	369	31	35	91.25	92.25	91.75
8	371	370	30	29	92.75	92.50	92.63
9	367	376	24	33	91.75	94.00	92.88
10	366	369	31	34	91.50	92.25	91.88
11	362	366	34	38	90.50	91.50	91.00
12	364	373	27	36	91.00	93.25	92.13
13	355	370	30	45	88.75	92.50	90.63
14	360	383	17	40	90.00	95.75	92.88
15	367	381	19	33	91.75	95.25	93.50
Total	5454	5621	379	546	90.90	93.68	92.29

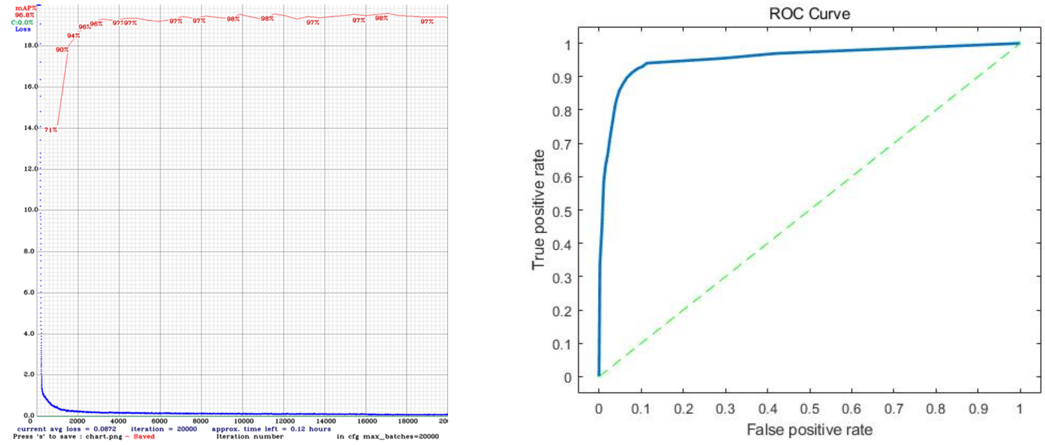
Note: TP: true positive; TN: true negative; FP: false positive; FN: false negative; Sen: sensitivity; Spe: specificity; Acc: accuracy.

test dataset is 51.42%. The details are shown in Figure 2.6.

2.3.3 Performance Evaluation

We evaluated the model performance by means of sensitivity, specificity, and accuracy. We checked the results of the three spinal endoscopic experts, and divided them into true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP refers to the images with nerve or dura mater that were correctly recognized and labeled by the trained model; TN refers to the images without nerve or dura mater that were not labeled by the trained model; FP refers to the images without nerve or dura mater that were misrecognized and mislabeled by the trained model; and FN refers to the images with nerve or dura mater that were misidentified and not labeled by the trained model.

In order to evaluate the stability of the model more objectively, we trained the model for 5 times. Then select the best weight to test the same test dataset, and calculate these evaluation indexes. Finally, we use confidence level and confidence



(a) Verification accuracy and loss curves

(b) ROC curve

Figure 2.6: The performance of the model in the process of training and testing.

interval (CI) to describe the performance of the model. The following formulas were used:

Sensitivity = TP / (TP+FN);

Specificity = TN / (TN+FP);

Accuracy = (TP+TN) / (TN+TP+FN+FP).

2.3.4 Test Results for CADS

After training the model once, we record the test results in Table 2.2. Then, we mix the original training and test data and shuffle the training and test datasets.

We randomly divided 6000 images into the test dataset, and the remaining images were used to retrain the model. Repeat 4 times, and the results are shown in Table 2.3. At the 95% confidence level, the CI of sensitivity was 90.62% - 91.17%, the CI of specificity was 93.38% - 93.97%, and the CI of accuracy was 92.23% - 92.35%. The results also show that the performance of the model is stable and reliable.

Table 2.3: Results of five tests on CADS.

Times	TP	TN	Sen(%)	Spe(%)	Acc(%)
1	5454	5621	90.90	93.68	92.29
2	5435	5633	90.58	93.88	92.23
3	5469	5604	91.15	93.40	92.28
4	5448	5636	90.80	93.93	92.37
5	5463	5609	91.05	93.48	92.27
95%CI	[5437,5470]	[5603,5638]	[90.62,91.17]	[93.38,93.97]	[92.23,92.35]

Note: CI: confidence interval.

2.3.5 Test Results for Four Surgeons

To analyze the results between CADS and the level of experience of doctors, four surgeons with different levels of experience examined the same images in test dataset.

Doctor #1 (D1) is a professor with more than 20 years of experience in spinal endoscopic operations and has performed over 150 PTEDs. Doctor #2 (D2) is an associate chief physician with more than 16 years of experience in spinal endoscopic operations who performs PTEDs independently. Doctor #3 (D3) is an associate chief physician of a primary hospital. He has worked in general surgery for 10 years, but seldom participates in spinal endoscopic surgeries and does not perform PTEDs independently. Doctor #4 (D4) is an attending doctor of spine surgery. He has participated in over 30 PTEDs as an assistant but has not operated independently.

We analyzed the sensitivity, specificity, and accuracy of the four doctors, and contrast their results with CADS. The results are shown in Table 2.4 and Figure 2.7.

2.3.6 Statistical Analysis

We used the chi-squared test to analyze the indicators in Table 2.2. The p values for each patient in test dataset were: sensitivity, $p = 0.462$ and specificity, $p = 0.1$. All of the p values were >0.05, meaning that there was no significant

Table 2.4: Comparison of four doctors and CADS

	TP	TN	Sen (%)	Spe (%)	Acc (%)
CADS	5454	5621	90.90	93.68	92.29
D1	5904	5936	98.40	98.93	98.67
D2	5796	5883	96.60	98.05	97.33
D3	5011	5107	83.52	85.17	84.32
D4	5309	5571	86.32	92.85	89.58
D1 combined with CADS	5919	5941	98.65	99.02	98.83
D2 combined with CADS	5833	5901	97.22	98.35	97.78
D3 combined with CADS	5539	5733	92.32	95.55	93.93
D4 combined with CADS	5651	5808	94.18	96.80	95.49

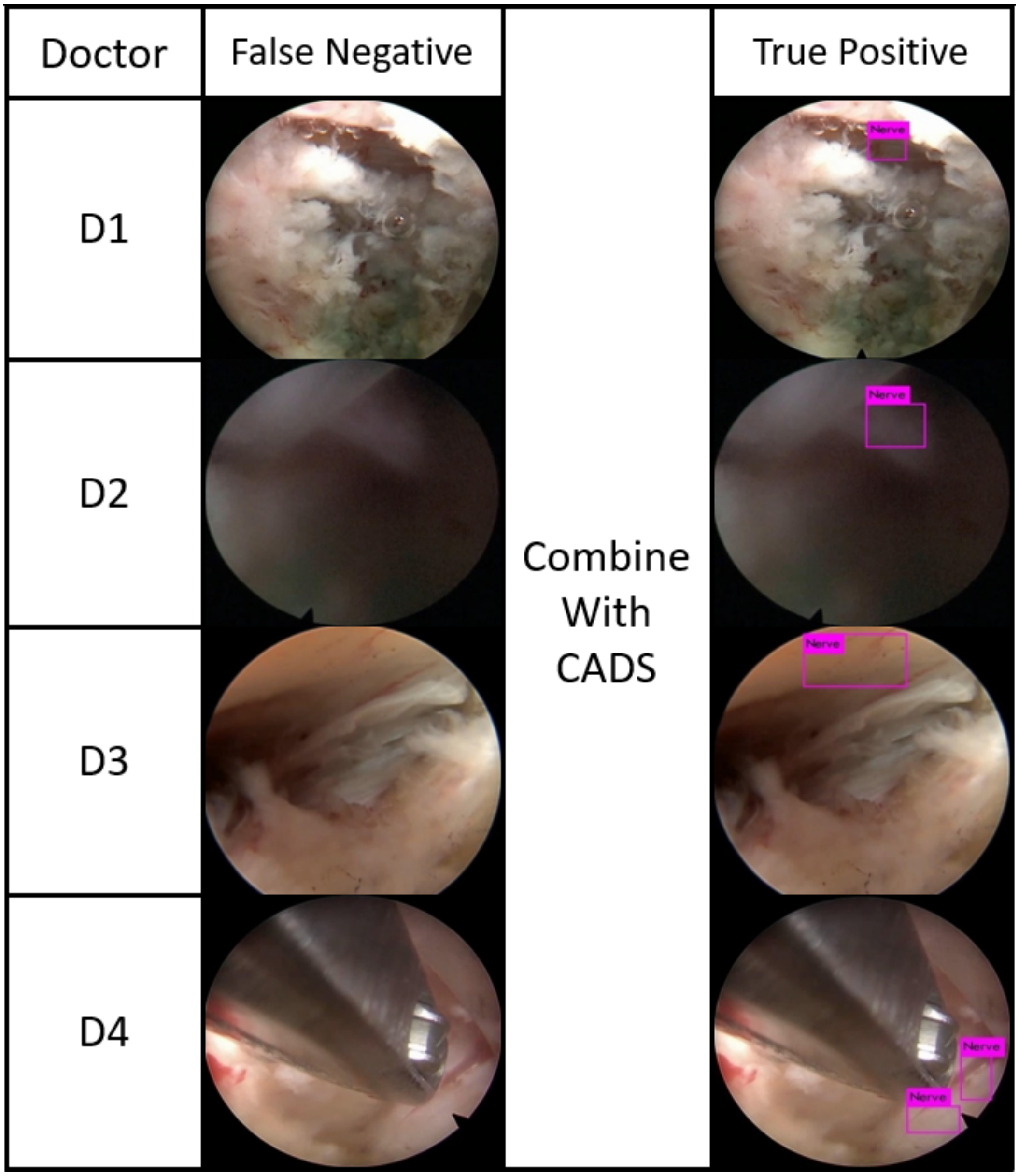


Figure 2.7: The result for combination of Doctors with CADs.

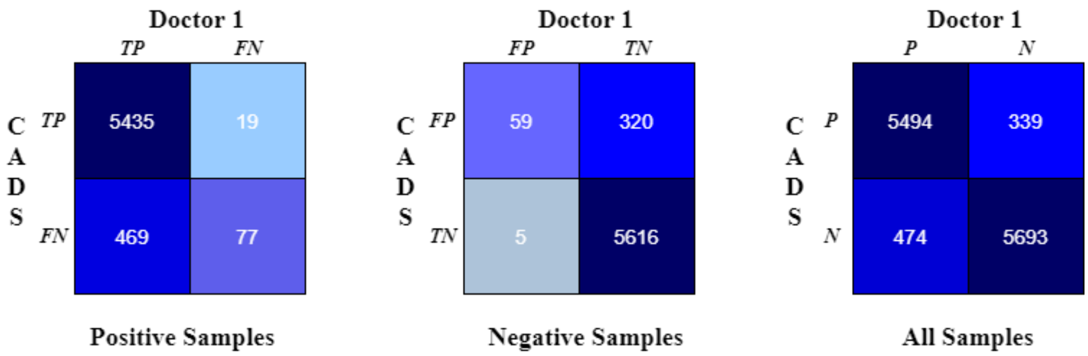


Figure 2.8: The confusion matrix for Doctor 1 and CADS.

difference in the sensitivity and specificity for recognizing nerve and dura mater in each patient.

We used paired t-test to analyze the indicators in Table 2.4. The p values of D1 and D1 combined with CADS were: sensitivity, $p = 0.055$ and specificity, $p = 0.136$, p values were greater than 0.05, which means that there was no significant difference in the sensitivity and specificity of D1 and D1 combined with CADS. The p values of D2 and D2 combined with CADS were: sensitivity, $p = 0.001$ and specificity, $p = 0.018$. For D3, D4, we also made the same statistical analysis, p values were lower than 0.001, which shows that D2, D3, D4 in combination with CADS, the indicators have been significantly improved, and the difference is significant.

We also use *kappa* coefficient to evaluate the consistency between the detection results of CADS and D1, and the confusion matrix is shown in Figure 2.8.

According to the calculation formula, we calculated the kappa coefficient of doctor 1 and CADS on the whole dataset, $k = 0.86$ (> 0.8), which indicates that the performance of CADS are highly consistent with the D1.

2.4 Discussion

Although CNN technology has been widely used in the recognition and segmentation of medical images, [74, 75] the recognition of spinal endoscopic images has not been investigated yet. By applying CNN-based methods, we identified the nerve and dura mater images under the spinal endoscope satisfactorily.

2.4.1 Image Features of Nerve and Dura Mater

In the process of spinal endoscopic surgery, spinal endoscopy experts noticed that in different states, the characteristics of nerve and dura mater are different. For example, when the nerve and dura mater are compressed, the surface of nerve and dura mater is pale. When there is inflammation, the capillaries on the surface of nerve and dura mater will expand, and the color is ruddy.

In this study, to let the CNN learn the characteristics of nerve and dura mater in different situations, we collected a number of PTED videos in varied environments. To reduce the occurrence of bias in the training and testing processes, we also focused on the balance of the quantity and quality of training images. Regarding the images of the 50 patients used for the current research, we took about 200 clear and unclear images of nerve and dura mater from the images of each patient's operation for labeling and training.

We also noticed the effect of light on the feature recognition of tissues. In the verification of the test results, we found that some images could not be correctly recognized and marked by experts and computer-aided diagnosis system due to the strong endoscope light. It is also one of our future research topics to compare the effects of doctors and CADS on tissue recognition under different light conditions by adjusting light intensity or spectrum.

2.4.2 Performance the CADS

From the training process and test results, it can be seen that there is an acceptable result of IoU in the training process, and the loss curve converges rapidly and tends to be stable within 1,000 iterations. The best average precision (AP) during the training process can reach 98% and keep at about 97%.

In the process of verifying the images labeled by CADS one by one, doctors were satisfied with the work of CADS. Although the higher the value of IoU, the better the performance of the model, but doctors believe that even if only part of the nerve or dura mater in the image is correctly labeled by CADS, it is enough to attract their attention. Surgeons will quickly find the areas where nerve and dura mater are located according to the bounding box to avoid injury. However, if CADS is to be used in the teaching of medical students or junior doctors to help them understand the features of nerve and dura mater, higher IoU will provide more accurate assistance.

To estimate the performance of CADS, we calculated the sensitivity, specificity, and accuracy, and evaluated these indicators for each patient in test dataset. If CADS has a large detection deviation in the recognition of nerve and dura mater images in the test dataset, the recognition accuracy of the model may be very high for some patients' nerve and dura mater images, while that for others' images may be very low. In that way, if CADS is used for recognition in surgery, for those patients with low recognition accuracy, because of the high rate of false recognition of images, it may lead to serious complications.

The results showed that there was no statistical difference in sensitivity, specificity, and accuracy of the model for each patient, meaning that the performance of the CADS was stable and good. On the other hand, although the age distribution of patients in our dataset is extensive, we compared the nerve and dura mater recognition indexes of different age groups by statistical methods. We found that there was no significant difference in the recognition rate of nerve and dura mater in patients of different age groups by CADS.

The CADS can achieve good recognition accuracy in a short training time and can maintain a high level of accuracy, which showed that the CADS has good learning ability. However, in the test dataset, we found that the accuracy of the CADS for object recognition is lower than that of the training process. We thought that one of the reasons for this situation is that the absolute number of training data is insufficient.

As the research goal, we also expect CADS to have higher precision and faster speed in the recognition of nerve and dura mater images under PTED. In order to make it useful in clinical practice, we need to improve the structure and parameters of the CNN, strengthen the computer hardware configuration, so as to improve its overall ability in real-time application.

2.4.3 Comparison of Performance among Surgeons with Different Levels of Experience

Four surgeons with different levels of experience identified the images in test dataset. The performance of D1 was significantly better than that of D2, CADS, D3, and D4 ($p<0.001$). However, when combining CADS with D3 and D4, sensitivity and specificity increased to (D3) 92.32% and 95.55% and (D4) 94.18% and 96.80%, respectively, and accuracy was (D3) 93.93% and (D4) 95.49%, values that were significantly higher than that of D3 and D4 alone ($p<0.001$), and approaching the performance of D2.

By analyzing the test results, it is not difficult to find that there is no significant difference ($p>0.05$) between the results of D1 and D1 combined with CADS. These results also indicate that CADS can significantly improve the recognition ability of D2, D3 and D4 to nerve and dura mater under PTED ($p <0.001$), but it cannot effectively improve the recognition ability of D1 ($p >0.05$); when D3 and D4 were combined with CADS, the performance still could not reach D2 ($p <0.001$).

At this stage, the application of CADS in teaching medical students or junior doctors may be an effective method, but its clinical application seems not mature enough yet.

2.4.4 Misrecognized and Unrecognized Images

Through analysis of unrecognized nerve and dura mater images, we found that when the nerve and dura mater areas in the image were <5% and operation vision was blurred, recognition accuracy is significantly reduced. Because these blurred images occur during intraoperative bleeding and irrigation, this is also a high incidence period of adverse complications in endoscopic surgery. In these conditions, experts recommend that surgeons do not perform any operation that may cause damage before the visual field is clear. This principle should also be followed in endoscopic surgery. However, if the CADS can recognize the nerve and dura mater images in a blurry environment more accurately, it will clearly make a valuable contribution to spinal endoscopic surgery.

We also found that the CADS has a very significant advantage in detection speed. For 12,000 images in the test dataset, CADS took 172 seconds to complete the detection; for four doctors, it took four to five hours. When we considering a long-time operation or continuous operation, CADS performance is not affected by time factor, and can maintain a stable recognition accuracy over a long-time period. However, doctors may gradually fatigue due to long-time concentration, and need to pay more attention to the operation to avoid damage to important tissues.

2.4.5 Limitations

Our study had some limitations. Firstly, all the patients and surgical videos came from one center using the same type of machine. Moreover, the number of images of nerve and dura mater examined here is likely insufficient. We should further increase the number of patients and surgical videos from different hospitals. Secondly, we only used one CNN model to train and test our dataset. An ensemble algorithm is a novel approach to blend multiple algorithms to improve predictive performance compared with any one algorithm alone, and it may be more effective when individual classifiers are not correlative; an ensemble algorithm works by removing uncorrelated errors of individual classifiers by averaging. [76–78] Thirdly, in the unrecognized images, we found that a few target images were clearly but not correctly identified by CADS, and when the target image was blurred or small, CADS did not show very high sensitivity. Finally, in the process of our research, we also tried to use CADS to test short video in real-time mode, but its performance is not as satisfactory as image recognition. We need to do more research in the future to solve these problems.

2.5 Summary

Artificial intelligence has been widely used in various medical fields, and has shown its potential in the past decade, providing doctors with convenient tools for diagnoses, treatment, and prognostic diseases. [79–81]

In this research, we used CADS based on CNN to recognize nerve and dura mater images in patients undergoing PTED. After we adjusted the model parameters and increased the target features, we achieved sensitivity, specificity, and accuracy of 90.90%, 93.68%, and 92.29%, respectively, with no significant difference between patients. We also compared the performance of four surgeons with different levels of experience with CADS. Although the performance of CADS was lower than that of the most experienced spinal endoscopist, it was higher than that of the general surgeons. Combined with CADS, the performance of the general surgeons approached that of the spinal endoscopist. This indicates that CNN can recognize the images of nerve and dura mater during PTED, and thus demonstrates an efficient method for tissue recognition that can assist general surgeons during PTED.

In future research, we aim to label more tissues in PTED surgical videos based on improving the accuracy of blurry image recognition. We also plan to use assemble CNN models to recognize these tissues more accurately and quickly, and realize real-time recognition of various tissues in the whole operation process. The goal is to help reduce the workload of spinal endoscopists and improve the performance of general surgeons, optimize the learning curve, and significantly reduce complications and improve the safety of operations.

Chapter 3

Comparison of eight CNN Models in Detecting Nerve and Dura Mater under PTED

3.1 Introduction

With the development of deep learning and the improvement of computing power, the computer vision community has witnessed significant progress in object detection. In the early 21st century, when the first CNN based object detection model R-CNN was proposed, the object detection model has been booming in various fields. Current state-of-the-art detectors [35–38, 44, 46, 82] show high performance on several very challenging benchmarks such as COCO [83] and Open Images [84].

In medical field, CNNs are not a solution to replace doctors, but will assist doctors to optimize their routine tasks, thus having a potentially positive impact on medical practice. [85] For the detection of target tissue under endoscopy, the current research focuses on the detection of target tissue under upper gastrointestinal endoscopy, colonoscopy, and laryngoscope, which has achieved good results and is gradually carried out in clinical practice [86–88]. However, the tissue recognition in endoscopic surgery has not been widely carried out.

In Chapter 2, we have used Yolov3 to recognize the nerve and dura mater under PTED, and achieved good performance. In this Chapter, We will briefly introduce some representative common object detection models, use them to learn the features of nerve and dura under PTED, and test their performance in recognition of nerve and dura mater under PTED.

These detectors can be divided into two categories, one-stage detectors and two-stage ones. One-stage detectors are more efficient and elegant in design, but currently the two-stage detectors have superiority in accuracy.

The following are some common CNN models for image recognition and target detection.

1. R-CNN

R-CNN is a region based CNN detector. It consists of four modules. The first module generates category independent region suggestions. The second module extracts a fixed length feature vector from each region. The third module is a set of class specific linear SVM, which are used to classify the objects in one image. The last module is a boundary box regression for accurate boundary box prediction.

2. Faster R-CNN

Faster R-CNN is to input the whole image into CNN for feature extraction; use RPN (region proposal network) to generate proposal windows, and 300 proposal windows are generated for each image, and then the proposal windows are mapped to the last convolution feature map of CNN; Then through the ROI pooling layer, each ROI generates a fixed size feature map; Finally, Softmax loss and Smooth L1 loss are used to train classification probability and bounding box regression.

3. Yolo

Yolo as a series of algorithms, applies a single neural network to the whole image, and then divides the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are weighted by the predicted probabilities. yolo requires only one forward propagation pass through the neural network to make predictions.

4. SSD

SSD extracts feature information from different convolution layers for detection. Different convolution layers have a variety of receptive fields and semantic information. Convolution layer is used to replace full connection layer and reshape layer, and the number of convolution cores is used to control the number of channels of output feature map. At the same time, prior anchor box is used, and each box is predicted.

5. RetinaNet

RetinaNet is a single, unified network composed of a backbone network and two task-specific subnetworks. The backbone is responsible for computing a convolutional feature map over an entire input image and is an off-the-self convolutional network. The first subnet performs convolutional object classification on the backbone's output; the second subnet performs convolutional bounding box regression.

6. Yolov3

Yolov3 uses Darknet-53 as the backbone, and uses residual network to increase the depth of the network and improve the accuracy. The internal residual block uses jump connection, which alleviates the gradient disappearance problem caused by the increase of depth in the deep neural network.

7. CenterNet

Centernet applies cascade corner pooling and center pooling in convolutional backbone network to output two corner heatmaps and a center keypoint heatmap, respectively. A pair of detected corners and the similar embeddings are used to detect a potential bounding box, and the center point of the boundary box is used to represent the target object. Then other attributes of the object are regressed based on the point, and the final boundary box is determined by using the detected center key points.

8. EfficientDet

EfficientDet uses residual neural network to increase the depth of neural network, and uses deeper neural network to realize feature extraction; By changing the number of feature layers extracted from each layer, more layers of feature extraction can be achieved, more features can be obtained, and the width can be improved. By increasing the resolution of the input image, the network can learn and express more abundant things, which is conducive to improving the accuracy.

9. Yolov4

Yolov4 is the improve version of Yolov3. It changed the backbone to CSPDarknet-53, uses SPP (Spatial Pyramid Pooling) and PANet (Path Aggregation Network) to strengthen feature extract, and use some tricks to improve the detection efficiency of the model.

10. Yolov4-Tiny

Yolov4-Tiny is a simplified version of Yolov4. In order to improve the detection speed, Yolov4-Tiny reduces the structure and only uses two feature layers for classification and regression prediction. The parameters are only six million, only one tenth of Yolov4.

The innovations of this study are as follows: (1) Eight popular CNNs were used to provide a comprehensive view of the role of artificial intelligence in recognize the nerve and dura mater images in PTED; (2) The detection performance of these CNN models were compared by sensitivity, specificity, accuracy, positive predictive value, negative predictive value, area under curve value, and detection speed; (3) We confirmed that CNN model is an effective tool for detecting nerve and dura mater in PTED, which can help surgeons identify nerve and dura mater accurately and timely.

3.2 Methods and Materials

This research has been approved by the Ethics Committee at Zhongnan Hospital, Wuhan University, Hubei Province, China.

All of these video data were collected from the department of orthopedics, Zhongnan hospital, Wuhan University. All the videos were PTED surgery videos carried out from Jan 1, 2019 to Dec 31, 2020 who underwent the PTED surgery with lumber intervertebral disc herniation. The inclusion and exclusion criteria were the same as our previous studies.

To protect the patient's privacy, we only collected the age, gender, illness, and surgical location of the patient and the video of the PTED as the research data. All of these endoscopic surgical were performed by senior spinal endoscopic experts of Zhongnan hospital.

3.2.1 Datasets

3.2.1.1 Information of Patients

In this study, total 102 patients' data were selected. 63 males and 39 females, ranging in age from 20 to 71 years old (56.15±7.42), all of these patients were having different level lumbocrural pain and limb numbness, 6 patients had LDH at two different sites. Detail information is shown in Table 3.1.

3.2.1.2 Training Dataset and Test Dataset

First of all, three spinal experts went through all the PTED videos in dataset, and marked the time when the nerve and dura mater appear in the videos, then we cut out these video clips and turned them into images. Three spinal experts checked and labeled all the nerve and dura mater images by LabelImg software. After three experts confirmed all the labels and reached a consensus, then, these images and annotations information were conserved according to the patients' name separately.

In order to keep balance in the training and test datasets, we built a database with 40,800 images (for each patient, we randomly selected 200 images with nerve and dura mater, 200 images without nerve and dura mater).

According to the ratio 8:2, we random select 81 patients' data (16,200 images

Table 3.1: Patient information

Dataset	Age (years)	Gender		Site of Lesion			Number of Images	
	Mean ± SD	Male	Female	L3/4	L4/5	L5/S1	Positive	Negative
Training	50.05 ± 13.81	48	33	13	36	18	16,200	-
Test	54.13 ± 10.27	14	7	5	14	5	4,200	4,200

with nerve and dura mater) for training dataset, and 21 patients' data (4,200 images with nerve and dura mater and 4,200 images without nerve and dura mater) for test dataset. In the training process, we random divided the training data into training and interval validation data in a ratio 9:1.

3.2.2 CNN Models and Training Parameters

In this study, 8 well-known pre-trained CNN model were used to recognition Nerve and dura mater from PTED. 1-CenterNet, 2-EfficientDet, 3-Faster-RCNN, 4-RetinaNet, 5-SSD (Single Shot MultiBox Detector), 6-Yolov3, 7-Yolov4, 8-Yolov4-Tiny.

3.2.3 Computer Configurations

The computer comprised an Intel i9 9900k CPU @3.60GHz × 16, RAM 32 GB, and a CUDA-enabled Nvidia Titan 312 GB graphics processing unit (Nvidia), based on the hardware of the NVIDIA GeForce RTX 2070 SUPER GPU.

We used transfer learning technology for download the pre-trained weight to optimize the CNN models' performance.

Sample images of the original data and the images of the labeled data are shows in Figure 3.1, the flowchart of this study is shown in the Figure 3.2. .

3.3 Results

In this study, the positive and negative cases were assigned to ND (nerve and dura mater) and non-ND (without nerve and dura mater), respectively. TP and

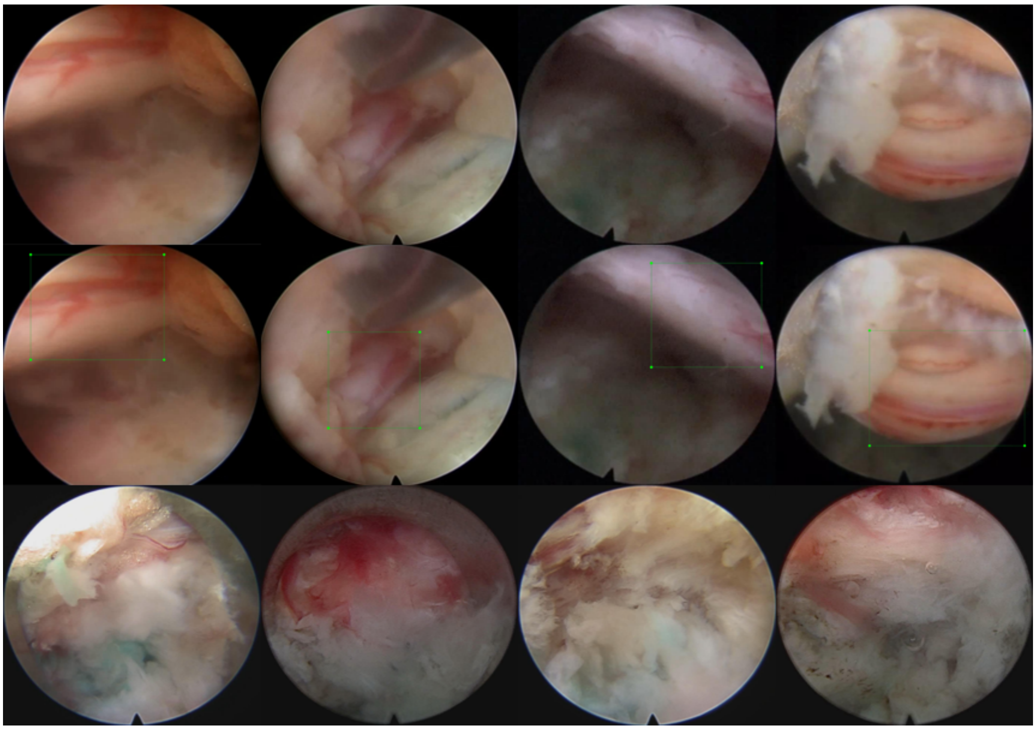


Figure 3.1: The positive samples(upper), labeled samples (middle), and negative samples (lower).

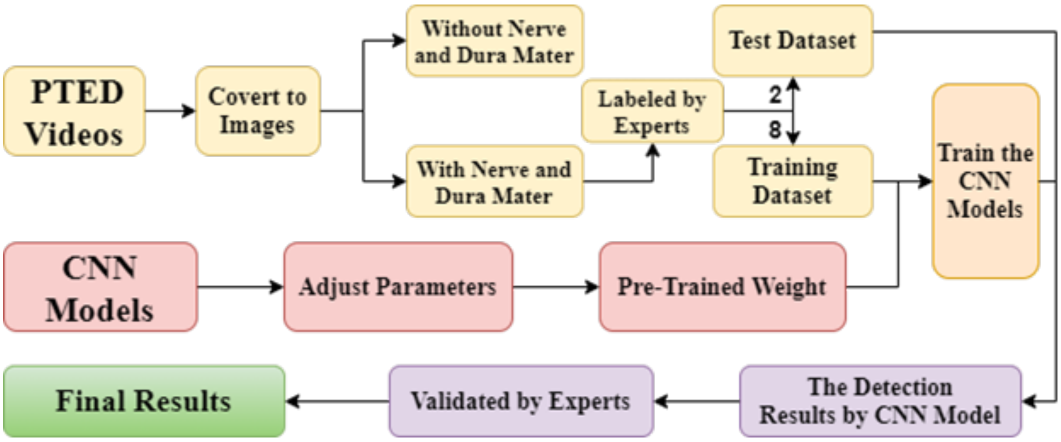


Figure 3.2: The flowchart of this study.

Table 3.2: The test results by different CNN models

CNN Models	True Positive	Ture Negative	False Positive	False Negative
CenterNet	4099	2667	1533	101
EfficientDet	4157	1155	3045	43
Faster-RCNN	4076	3189	1011	124
Retinanet	4100	2569	1631	100
SSD	4094	3666	534	106
Yolov3	3902	3943	257	298
Yolov4	3893	4038	162	307
Yolov4-Tiny	3882	3609	591	318

Note: All the results are based on confidence level 0.1.

TN represent the number of correctly labeled on ND and non-labeled on non-ND, respectively. FP and FN represent the number of incorrectly labeled on non-ND and non-labeled on ND, respectively. The purpose of our study is to understand the ability of CNN models to recognize nerve and dura mater in PTED. Therefore, we show the test results of all models at the confidence level 0.1 and the results are shown in table 3.2.

We use the following indicators to evaluate the performance of these CNN models: Sensitivity (Sen), Specificity (Spe), accuracy (ACC), positive predictive value (PPV), Negative predictive value (NPV). We also plot the receiver operating characteristic (ROC) curve and AUC value as indicators to evaluate the performance of these models. In order to find the best performance on each CNN model, Youden index was used to find the optimal threshold for each model, and the best performance of each model are shown in Table 3.3 and the Figure 3.3. The calculate formulas are as follow and the results are shown in Table 3.3, the ROC curve is shown in Figure 3.3.

Sen = TP / (TP + FN);

Spe = TN / (TN + FP);

PPV = TP / (TP + FP);

NPV = TN / (TN + FN);

Acc = TP + TN / (TP + TN + FP + FN).

Table 3.3: Performance comparison of eight models using test dataset

CNN Models	Sen(%)	Spe(%)	PPV(%)	NPV (%)	Acc (%)	FPS	AUC
CenterNet	91.02	84.24	85.24	90.37	87.63	75.78	0.926
EfficientDet	79.00	77.17	77.58	78.61	78.08	37.72	0.862
Faster-RCNN	82.43	82.67	82.63	82.47	82.55	15.62	0.920
Retinanet	90.43	86.17	86.73	90.00	88.30	33.59	0.926
SSD	92.43	94.21	94.11	92.56	93.32	44.31	0.972
Yolov3	92.90	93.88	93.82	92.97	93.39	36.31	0.945
Yolov4	92.69	96.14	96.00	92.93	94.42	28.12	0.952
Yolov4-Tiny	86.52	91.86	91.40	87.21	89.19	196.15	0.941

Note: TP: true positive; TN: true negative; FP: false positive; FN: false negative.

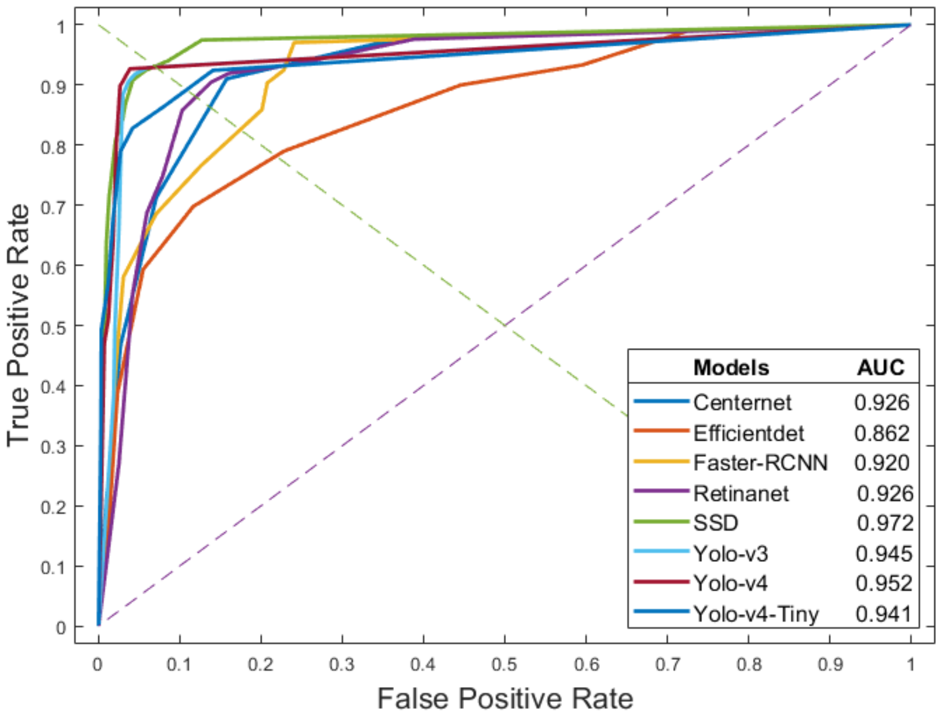


Figure 3.3: The ROC curves of eight networks.

The results in Table 3.2 show that EfficientDet has the best detection performance on true positive. For 4200 positive samples, 4157 images can be correctly detected. Yolov4 has the best detection performance on true negative. For 4200 negative samples, only 162 images were mistaken for positive. It is observed that all CNN models have high sensitivity to the recognition of nerve and dura mater, but there are great differences in specificity. The Youden index can help to find the best threshold to balance sensitivity and specificity. The best performance of each model is shown in Table 3.3.

For real-time object detection, higher speed models have the advantage for practical operation. In this study, based on our test dataset, Yolov4-Tiny model is the fastest model, detection speed can reach 196.15 frames per second. The video used in this study was 25 fps.

We chose the same images that recognized by all CNN models, and contrast the difference on confidence level, the images are shown in Figure 3.4 and Figure 3.5.

3.4 Discussion

In this study, eight common CNN models were used to provide a comprehensive test, using CNN technology to identify nerve and dura mater images under PTED. The results show that CNN can accurately identify nerve and dura mater from other tissues.

3.4.1 Detection Precision

According to the results show in Table 3.2, the highest sensitivity is produced by EfficientDet, for all positive samples, sensitivity can reach 98.98%, but at the same confidence level, the sensitivity is only 27.50%, it means a large number of false positive results is produced.

AUC value was used to evaluate the overall performance of these models ob-

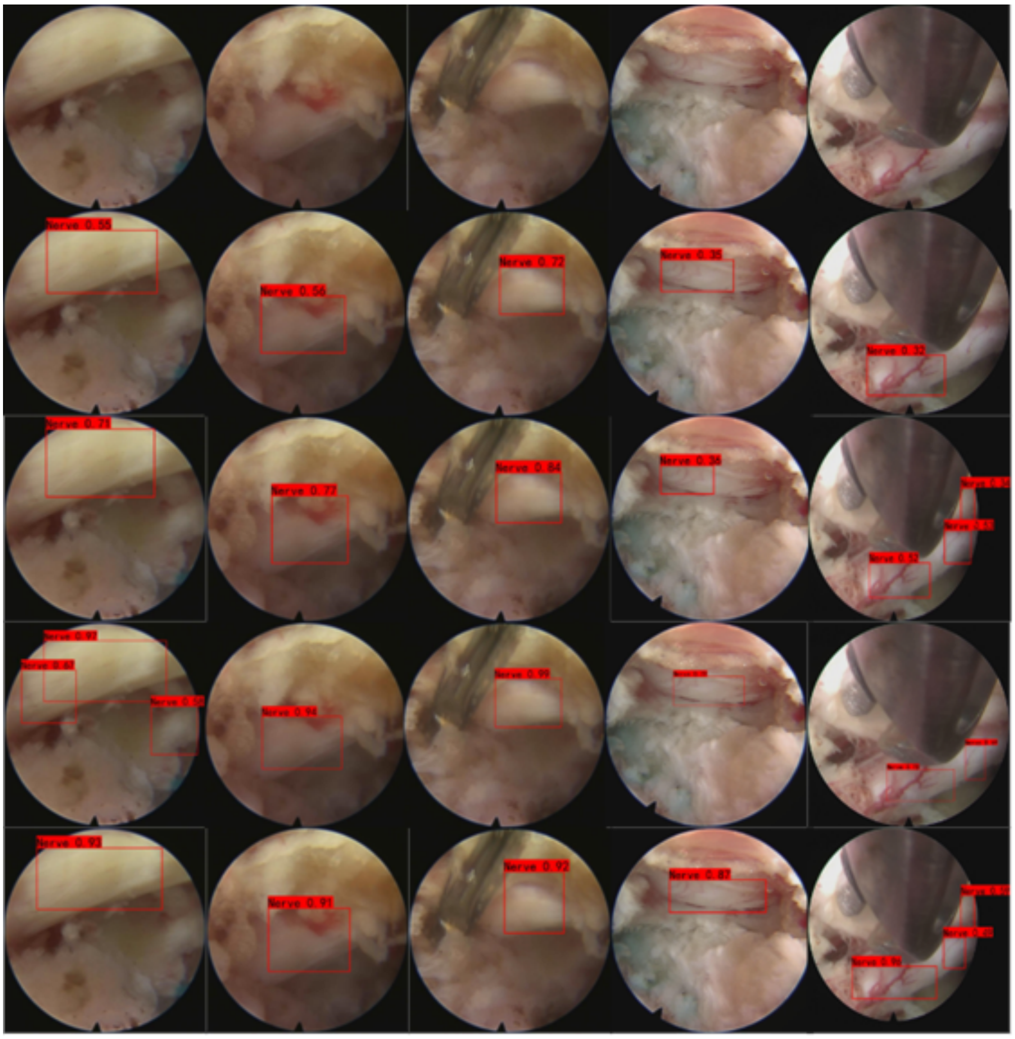


Figure 3.4: The part of test results detected by the different CNN models with the same images. The first row shows the original images in the test dataset, 5 different images as the samples to show the performance of each model. From row 2 to row 5 are the results detected by Centernet, Efficientdet, Faster-RCNN, Retinanet. The red rectangles are the detection box, the result of classification and confidence level are above the rectangles.

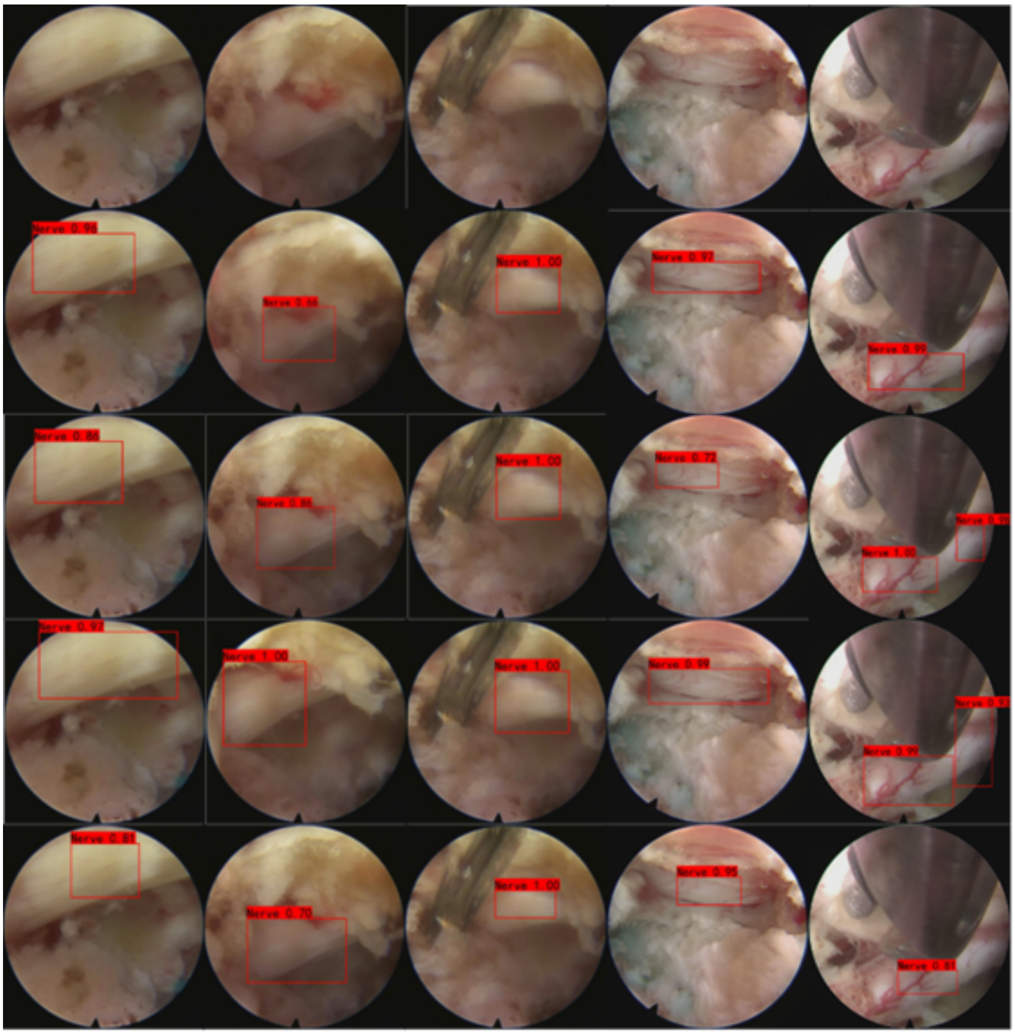


Figure 3.5: The first row shows the original images in the test dataset, 5 different images as the samples to show the performance of each model. From row 2 to row 5 are the results detected by SSD, Yolov3, Yolov4, Yoliv4-Tiny. The red rectangles are the detection box, the result of classification and confidence level are above the rectangles.

jectively. The best AUC value is produced by SSD 0.972. In Table 3.3, we can observe the performance of all CNN models clearly. Although the AUC value of the Yolov4 is lower than SSD, for the best performance, the Sen 92.69%, Spe 96.14%, PPV 96.00%, NPV 92.93%, Acc 94.42% are all higher than SSD. This also means that SSD still has good performance even when the confidence is very high.

We used Chi-square test to analyze the performance of SSD and Yolov4. For sensitivity, $p = 0.647$ (> 0.05), which means that there is no significant difference between SSD and Yolov4. For specificity, $p < 0.001$, for accuracy, $p = 0.003$ (< 0.05), which indicates that the specificity and accuracy of SSD and Yolov4 is statistically different. In the comprehensive comparison of various evaluation indicators, the overall performance of yolov4 is better than SSD.

Figure 3.4 and 3.5 showed some detection results. For the same positive images, Faster-RCNN, SSD, and Yolov4 have higher confidence level than other models.

3.4.2 Detection Speed

For detection speed, Yolov4-Tiny achieved the best performance, the FPS can reach 196.15. As we know, the number of parameter in Yolov4 is ten times than yolov4-Tiny. Actually, Yolov4-Tiny also have a high precision, the AUC is 0.941, the optimal threshold is 0.2, with sensitivity 86.52%, specificity 91.86%, and accuracy 89.19%.

All the video data in this study are 25 fps. As can be seen from Table 3.3, except for the two-stage model Faster-RCNN, the detection speed of all other models is higher than 25 fps, which indicates that all the one-stage CNN models we used have the ability to complete real-time object detection.

In this study, we used the two-stage model Faster-RCNN for the same test. However, the test shows that when the confidence is 0.1, the sensitivity of the model can reach 97.04%, but the specificity is only 75.93%. According to the Youden index, at the best detection point of the model, the sensitivity was 82.43%

and the specificity was 82.67%. Its detection speed is the slowest among all test models, and the detection speed is only 15 fps.

3.4.3 Limitations

Although most of the models we used can got satisfactory test results, the best results we get still can not meet the performance of spine surgeons who conduct similar tests in Chapter 2. This study also has the following limitations.

Firstly, the performance of the proposed eight CNN models was not objectively compared with that of clinical spine surgeons. So, future studies should plan to compare the CAD system with spinal surgeons at the same time.

Secondly, Compared with the performance tested in Chapter 2, the performance of the same CNN model Yolov3 is improved in this dataset. It seems that when we expand the training dataset, the performance of the model will still be improved. In the next research, we will start from expanding the amount of training data to further test the performance of the model.

Finally, this study did not test the performance of CNN models on the video dataset. Although the test results showed that the detection speed of all one-stage CNN models can exceed the frame rate of the video used in our test (25 fps), the detection results using real video data sets will be more convincing. Although this may cost more human and material resources, our research team will work together to solve this problem in the future.

3.5 Summary

In this chapter, eight CNN models use the same training dataset, the same annotation information and similar parameters to train each model. Then, we use the same test dataset to test the trained model, and use the same evaluation index to compare the performance of each model. The results show that all CNN models have good nerve and dura mater detection performance.

From the comprehensive model performance, SSD is the best one with a highest AUC 0.972. For the detection precision, the best performance is produced by Yolov4 with Sen 92.69%, Spe 96.14%, PPV 96.00%, NPV 92.93%, Acc 94.42% (confidence 0.1). For the detection speed, Yolov4-Tiny is the fastest model with 196.15 FPS. This study showed that the CNN models can be considered as a promising model to detect the nerve and dura mater under the PTED, and Yolov4 is the appropriate models to operate this task.

In fact, during the real surgery, surgeons can recognize the target tissues not only by the character of target tissues, but also refer to adjacent tissues and anatomical to avoid injury or protect the target tissues. However, since computerized image analysis can recognize the pattern quantitatively with higher precision, the CNNs could classify and detect with improved performance. The results of the present study indicate that the detection ability of the proposed CNNs are improved and all of them have high performance. Also, the deep learning technique can accurately aid surgeons to detect the nerve and dura mater during the PTED.

Normally, for operating the PTED surgery, requests for rich experience experts to operate the surgery, and the increased workload can affect the operation performance of experts and surgery effect. The results show that CNN can effectively assist surgeons to identify the images of nerve and dura mater, which is helpful to the development of endoscopic surgery, and reduce the complications of nerve and dura mater injury or tear.

Chapter 4

Vas Deferens Recognition under Laparoscopic Using CNN

4.1 Introduction

As we mentioned before, TAPP and TEP are the common invasive surgery for inguinal herniation. Although the development of TAPP and TEP has brought great convenience to patients, there are great challenges for doctors.

Laparoscopic minimally invasive surgery method can reduce the extent of skin incisions, nerve damage, and hematoma; lower postoperative pain and risk of infection of the surgical site; and lead to quicker recovery [89, 90], but it also has several disadvantages: for example, the surgeon initially needs a longer surgery time before plateauing on his/her learning curve; the surgery has a higher risk of complications; it needs much more knowledge of pelvic anatomy; and it needs a high level of surgical skill [91, 92], which can lead to more mistakes and harm to patients during the learning process.

In clinical practice, young surgeons need a long learning curve to carry out laparoscopic repair surgery well, especially TEP technology. With the increase of surgical experience and familiarity with local anatomy, complications and recurrence rates will gradually decrease [93].

Common complications of the operation include bleeding, bladder lesions, intestinal obstructions, intestinal perforations, injury to the iliac vein, femoral nerve, vas deferens, and even death [91].

With the development of artificial intelligence (AI) technology, CNNs have become an effective method for medical image analysis, disease prediction and diagnosis, and lesion detection and have been widely used [81, 94, 95]. CNNs are not a solution to replace doctors, but will help doctors optimize their routine tasks, thus having a potentially positive impact on medical practice [85].

There are many neural network models for object detection [38, 39, 96, 97], but each neural network has its own advantages. Some can detect objects quickly, but the precision is not optimal, while others can detect an object with higher precision, but the speed of detection is quite slow.

In the previous chapter, we have applied CNN to nerve and dura mater recognition under PTED and achieved satisfactory results. We also compared the performance of eight different CNN in this task. Yolov4, as an effective model, shows good performance in this task. Yolo is a one-stage convolutional neural network for object detection [98, 99], whose rate of detection can reach 65 fps with an average precision of up to 43.5% on the COCO dataset [99]. In this study, we propose to use Yolov4 as a detector to detect vas deferens images under TAPP and TEP.

The innovations of this study are as follows: (1) combined with computer CNN technology and clinical data, a new method for identifying vas deferens images in laparoscopic inguinal hernia repair using the CNN model is proposed, and the object detection ability of the CNN model (Yolov4) on the medical dataset is also tested; (2) different annotation methods are used to train the CNN model and to examine the performance of the model in the process of training and testing; (3) discussed with clinical experts and selected the appropriate IoU value to evaluate the performance of the model for reference by clinical surgeons.

4.2 Materials and Methods

This research was approved by the Ethics Committee at Hannan Hospital, Hannan District, Wuhan City, Hubei Province, China.

In this study, 35 adult male patients with inguinal hernia disease admitted to hospital for laparoscopic surgery from April 2018 to December 2019 were selected. The laparoscopic image device used was KARL STORZ-Endoscopy (22202020-110), America. All patients underwent laparoscopic hernia repair and signed a informed consent form. We collected information such as gender, age, disease name, and interoperation videos. All endoscopic surgeries were performed by senior endoscopic experts at Hannan Hospital. Details of the dataset are shown in Table 4.1.

Table 4.1: Dataset and patient information.

Dataset	Number of patients	Age(year)	Number of images
Training	26	63.15 ± 7.64	2,600
Image test	6	64 ± 8.12	1,200
Video test	3	55 ± 6.56	5,433

Note: All patients are male.

Inclusion Criteria

We selected those subjects satisfying all of the following three criteria: (i) adult male; (ii) the patient was diagnosed with inguinal hernia and underwent hernia repair for the first time; and (iii) the patient agreed to allow the use of video recordings for scientific research.

Exclusion Criteria

We excluded subjects matching any of the following three criteria: (i) female patient; (ii) patients with irreducible inguinal hernia; and (iii) the laparoscopic surgery was converted to open surgery for any reason.

4.2.1 Datasets

The patients were randomly divided into three groups, 26 patients in the training dataset, 6 patients in the image test dataset, and 3 patients in the video test dataset. In training dataset and image test dataset, we divided the surgical videos into images according to different datasets and saved them. A laparoscopic expert selected these images manually, then the other laparoscopic expert verified them, and finally deleted the disputed images and reached an agreement. In the video test dataset, we selected two short video clips from each patient's full surgical video, each of which is 30 seconds (30 frames per second). One of them is a clip with vas deferens image, and the other is a clip without vas deferens image. Then two laparoscopic experts verified these video clips and reached an agreement.

In order to balance the training data, we selected 100 images containing the vas deferens for each patient in the training dataset. In the image test dataset, we chose 200 images for each patient (100 images that included the vas deferens and 100 images without the vas deferens). Thus, a total of 3800 pictures of vas deferens and 180 seconds video clips were chosen to form an experimental database. There was no overlap in patients and images between training dataset, image test dataset and video test dataset.

All the training data were labeled using the software labelImg, and then validated by two laparoscopic experts; none of the researchers had any objection to the labeling results. The research flowchart is shown in Figure 4.1. The original images with vas deferens, labeled images, and the original images without vas deferens are depicted in Figure 4.2.

4.2.2 CNN Model and Training Parameters

In order to achieve higher accuracy with a faster speed, we used a one-stage neural network, Yolov4, to train and test the above dataset. The model structure is shown in Figure 4.3.

For training the model, we randomly divided the training dataset (2,600 im-

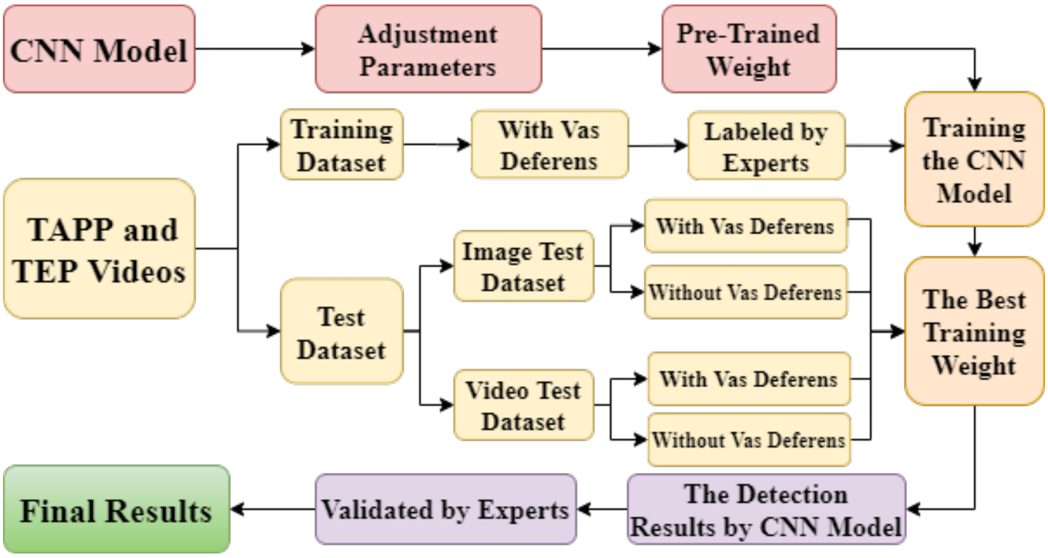


Figure 4.1: Research flowchart.

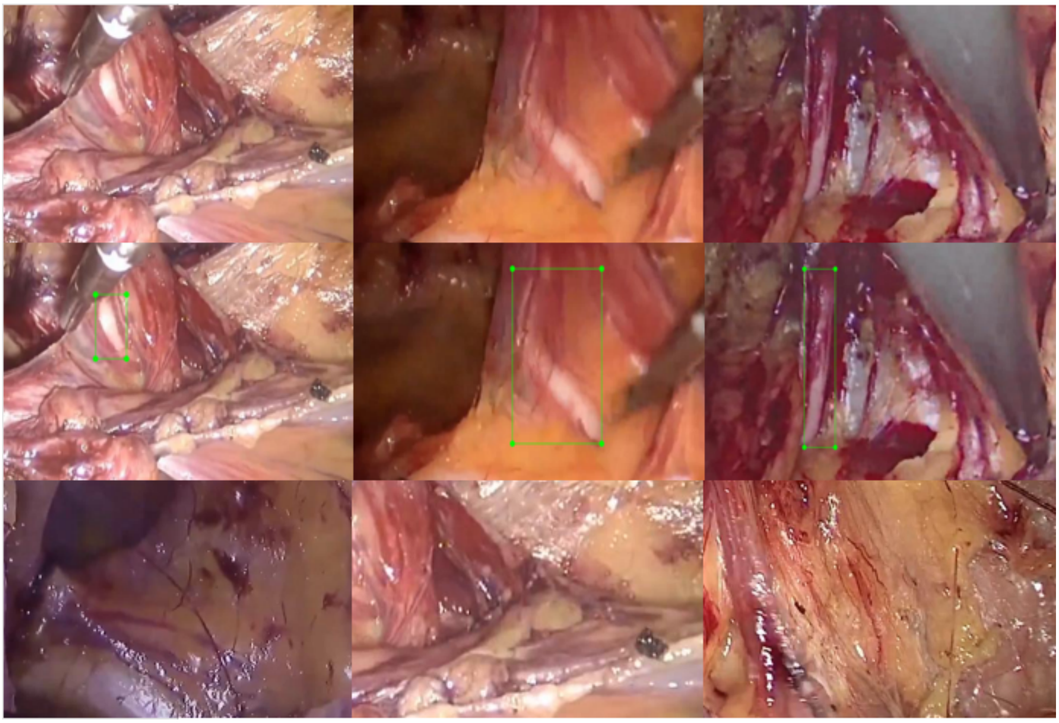


Figure 4.2: Original images. The first row shows original images with vas deferens; the second row shows labels included by the experts; the third row shows original images without vas deferens.

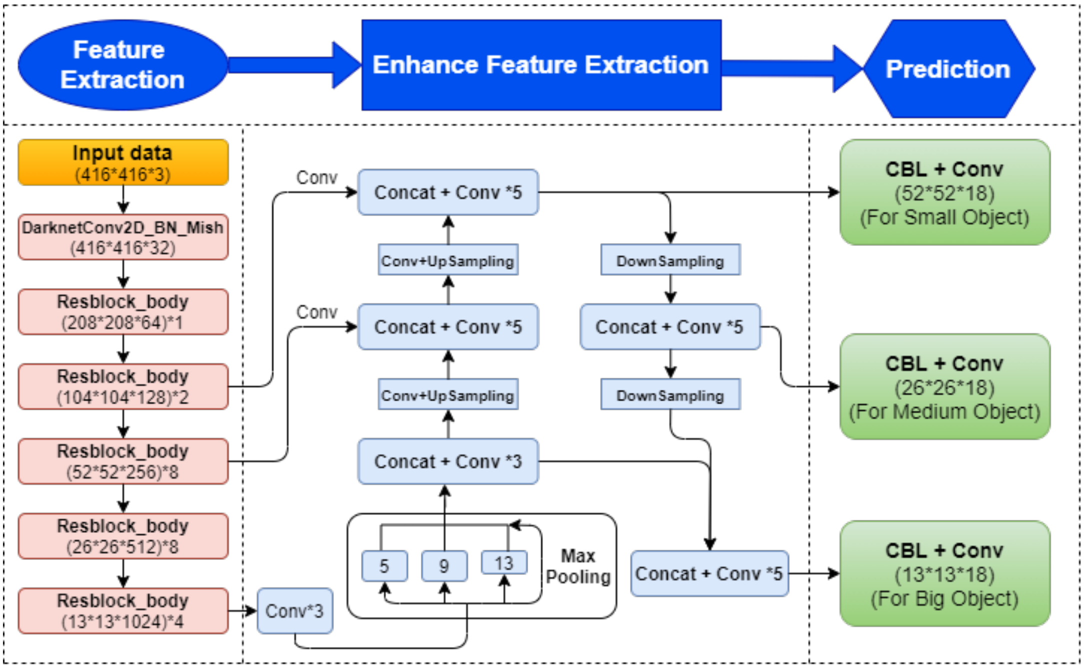

Feature Extraction
Enhance Feature Extraction
Prediction
Input data (416*416*3)
DarknetConv2D_BN_Mish (416*416*32)
Resblock_body (208*208*64)*1
Resblock_body (104*104*128)*2
Resblock_body (52*52*256)*8
Resblock_body (26*26*512)*8
Resblock_body (13*13*1024)*4
Conv*3
Conv
Conv
Concat + Conv *5
Conv+UpSampling
Concat + Conv *5
Conv+UpSampling
Concat + Conv *3
5
9
13
Max Pooling
DownSampling
Concat + Conv *5
DownSampling
Concat + Conv *5
CBL + Conv (52*52*18) (For Small Object)
CBL + Conv (26*26*18) (For Medium Object)
CBL + Conv (13*13*18) (For Big Object)

Figure 4.3: CNN Model Structure

ages) into training data and internal validation data according to the ratio of 9:1. Neural network parameters: input size = 416*416, batch = 64, subdivisions = 32, initial learning rate = 0.001, momentum = 0.95, max-batches = 10000.

Because our study is only a binary classification task, vas deferens tissue will only appear in one area of an image in laparoscopic IHR. Therefore, in the target detection stage, we adjusted the parameters in the process of non-maximum suppression (NMS) and set NMS-IoU to 0.1 to reduce the number of detection boxes.

The computer was an Intel i9 9900k CPU @3.6 GHz × 16, RAM 32GB, with a CUDA-enabled Nvidia Titan 312 GB graphics processing unit (Nvidia), based on the hardware of the NVIDIA GeForce RTX 2070 SUPER GPU. The whole training time was about 12 hours.

4.3 Results

We examined the test dataset using the best training weight in the training process, and then two laparoscopic experts verified whether the images in the test dataset were correctly labeled by the model. The two laparoscopic experts verified the labels in all the images and further discussed any labeled images that were controversial. Finally, they reached an agreement on all the labeling results.

For 1,200 images in the image test dataset, the model only takes 20 seconds to complete the target detection.

We defined those images with vas deferens that were correctly labeled as true positive (TP); the images without vas deferens but wrongly identified and labeled incorrectly as false positive (FP); the images containing vas deferens but not identified and not labeled, as false negative (FN); and the images without vas deferens and without any label as true negative (TN).

In the image test dataset, we used different confidence levels from 0.1 to 0.9 to evaluate the performance of the model. The model was used to label the images in the image test dataset, and two laparoscopic experts examined these results. For 600 positive symbols (images include vas deferens), a total of 607 detection boxes were labeled on these images. The detailed test results at different confidence levels are shown in Table 4.2. Example images of TP, FP, and FN are shown in Figure 4.4.

Table 4.2: Test result at different confidence levels.

Confidence Level	**TP**	**FP**	**TN**	**FN**
≥ 0.1	577	56	544	30
≥ 0.2	565	32	568	42
≥ 0.3	555	17	583	52
≥ 0.4	550	8	592	57
≥ 0.5	545	4	596	62
≥ 0.6	531	2	598	76
≥ 0.7	516	1	599	91
≥ 0.8	494	0	600	113
≥ 0.9	461	0	600	146

Note: These results are based on all the detetion boxes.

According to the test results in Table 4.2, we used these indicators to evaluate the performance of the CNN model: sensitivity (Sen), specificity (Spe), accuracy (Acc), positive predictive value (PPV), intersection over union (IoU), average

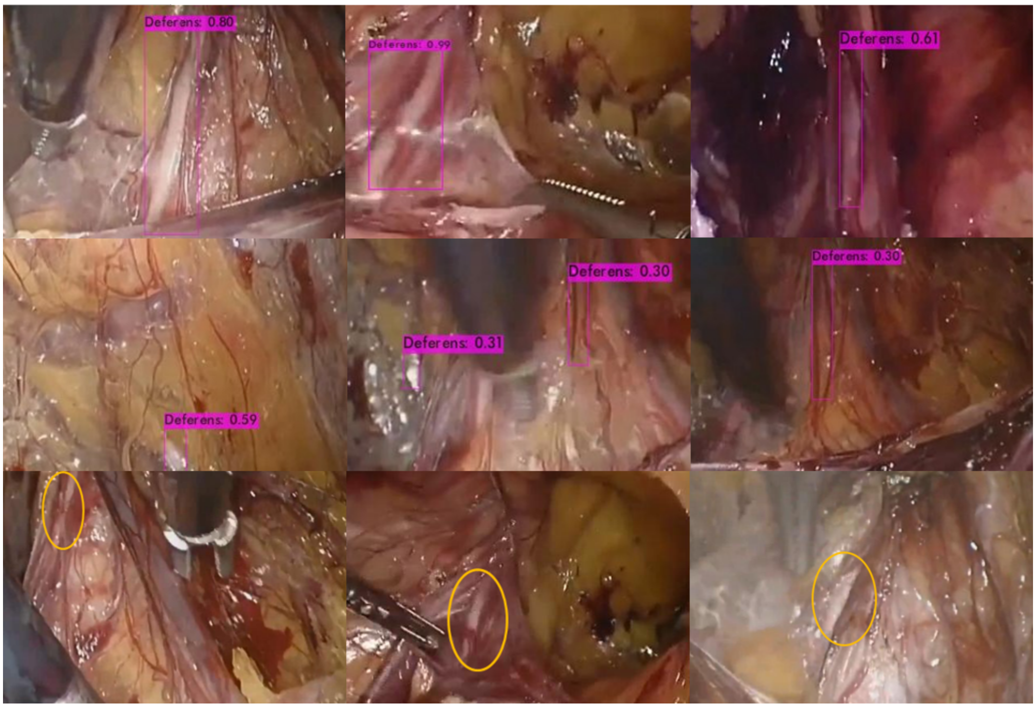


Figure 4.4: CNN-labeled vas deferens area and corresponding confidence level. The first row shows the true positive (TP); the vas deferens is labeled with a purple rectangle and the confidence level is shown above the rectangle. The second line shows false positive (FP). The third line shows false negative (FN). The model does not give any labels on the image. The researchers circled the area of the vas deferens in yellow.

precision (AP), and F1 score. We also draw the receiver operating characteristic (ROC) curve and calculate AUC value as indicators to evaluate the performance of the model. The formulas used to calculate these values are as follows and the results are shown in Table 4.3. The ROC curve is shown in Figure 4.5, The performance of the model for different IoUs are shown in Table 4.4.

Sen = TP / (TP + FN)

Spe = TN / (TN + FP)

PPV = TP / (TP + FP)

Acc = TP + TN / (TP + TN + FP + FN)

F1 = 2 * PPV * Sen / (PPV + Sen)

Table 4.3: Evaluation indicators at different confidence levels.

Confidence Level	**Sen(%)**	**Spe(%)**	**PPV(%)**	**Acc(%)**	**F1(%)**
≥ 0.1	95.06	90.67	91.15	92.87	93.06
≥ 0.2	93.08	94.67	94.48	93.87	93.85
≥ 0.3	91.43	97.17	97.03	94.28	94.16
≥ 0.4	90.61	98.67	98.57	94.61	94.42
≥ 0.5	89.79	99.27	99.27	94.53	94.29
≥ 0.6	87.48	99.62	99.62	93.54	93.16
≥ 0.7	85.01	99.81	99.81	92.38	91.82
≥ 0.8	81.38	100.00	100.00	90.64	89.73
≥ 0.9	75.95	100.00	100.00	87.90	86.33

According to the ROC curve, we calculate that the optimal confidence threshold of the model is around confidence level 0.4 and the AUC value is 0.9716, so in the process of testing the video test dataset, we use the confidence level of 0.4 to test the real-time detection function of the model. After saving the video detection results, we decompose the video clips into images for verification. A total of 5433 images were decomposed from the video clips (2719 with vas deferens and

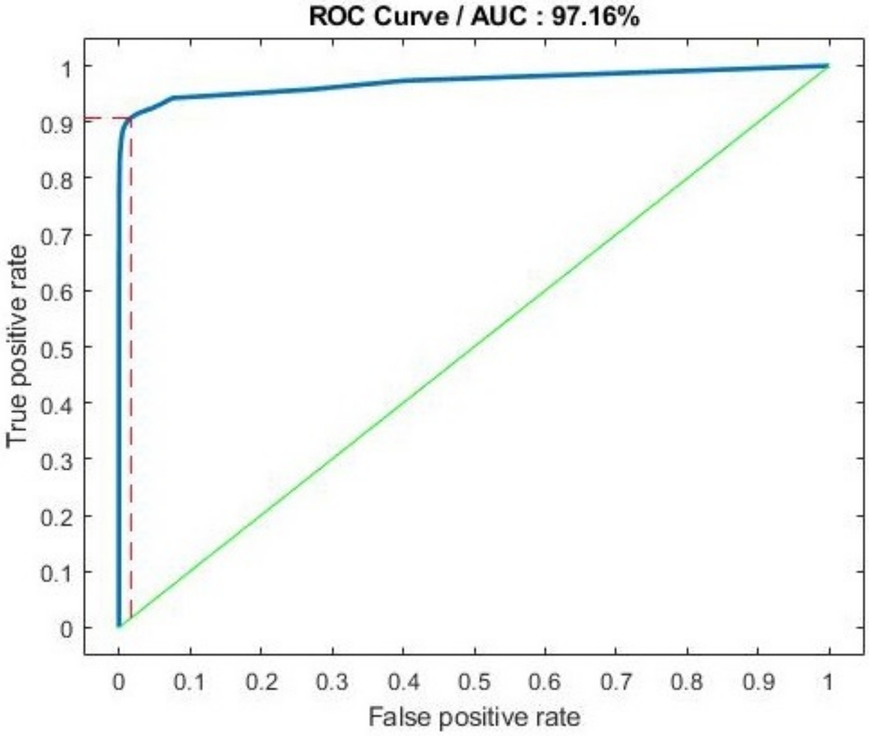


Figure 4.5: The ROC curve and AUC value.

Table 4.4: Evaluation indicators under different IoUs.

IoU	Sen(%)	PPV(%)	AP(%)	F1(%)
≥ 0.2	88.73	99.82	95.64	93.95
≥ 0.3	87.75	98.71	92.38	92.91
≥ 0.4	85.95	96.69	89.31	91.00
≥ 0.5	80.72	90.81	80.88	85.47
≥ 0.6	72.22	81.25	68.73	76.47
≥ 0.7	52.29	58.82	36.97	55.36

Note: These evaluation indicators are calculated based on 600 positive data items in the test dataset.

2714 without vas deferens). Two laparoscopic experts verified these images one by one and confirmed the results. The evaluation indicators and specific results are shown in Table 4.5.

Table 4.5: Test results in video test dataset.

Patients	Sen(%)	Spe(%)	PPV(%)	Acc(%)
1	90.23 (822/911)	96.14 (872/907)	95.92	93.18
2	89.50 (810/905)	95.23 (859/902)	94.96	92.36
3	90.58 (818/903)	95.91 (868/905)	95.67	93.25
Average	90.11 (2450/2719)	95.76 (2599/2714)	95.52	92.93

Note: These evaluation indicators are calculated based on confidence level 0.4.

In the video test dataset, these results were analyzed using IBM SPSS Statistics version 23.0. We used chi-square test to analyze the statistical differences between patients, with p >0.05 meaning no significant difference between the tested objects. The p value of sensitivity was 0.768, the p value of specificity was 0.608, all p values were greater than 0.05; the average Sen, Spe, PPV, Acc were 90.11%, 95.76%, 95.52%, 92.93% respectively. The results show that the model can effectively identify the vas deferens images in laparoscopic inguinal hernia repair, and the sensitivity and specificity of the model to different patients in the video test dataset are not statistically different.

4.4 Discussion

Using CNNs for medical image detection is not new, but using this technology to detect vas deferens under laparoscopic IHR surgery is novel; we use the CNN model to train and test the vas deferens images and obtained good results.

4.4.1 Performance of the CNN-based identification

In order to select the appropriate parameters and indicators to evaluate the performance of the model, we set different confidence levels to calculate the evaluation indicator. Laparoscopic experts verified the labeled images one by one and concurred on the final results. According to Table 4.3, it is clear that with increasing confidence levels, Sen gradually decreased from 95.06% to 75.95%, while Spe and PPV gradually increased from 90.67% and 91.15%, respectively, to 100%. In order to observe the performance of the model more comprehensively, we also calculated the Acc and F1 score, both of which first increased and then decreased with increasing confidence levels.

The F1 score is often used as a comprehensive index to judge the performance of a CNN model; it is a combination of Sen and PPV. When the confidence level is 0.2, 0.3, and 0.4, the F1 score of the CNN model is higher than 94%. ROC curve can show the influence of different thresholds on the generalization performance of the model, which is helpful to select the best threshold. [100, 101] We analyzed the ROC curve, compared the F1 score, and discussed with laparoscopy experts. Finally, we thought that when the confidence level was 0.4, the comprehensive performance of the model was more suitable for the dataset.

4.4.2 Medical Image Labeling Method

Using the correct labeling method is also an important aspect of the CNN model's target detection to achieve good results. Due to the complex environment of endoscopic surgery, target detection in endoscopic surgery is also a new challenge. There is currently no clear method for labeling the target tissue in endoscopic surgery.

At the beginning of this study, when we labeled the vas deferens on the surgical images, because the features of the target do not take on a regular shape, in order to lessen the proportion of non-vas deferens tissue in the label box and reduce the influence of non-vas deferens tissue on the target tissue, we only labeled the area with obvious vas deferens features on the image. However, the results show that the model is unable to obtain all information pertaining to the vas deferens in these images, which leads to unsatisfactory training and testing results.

We adjusted the labeling method and expanded the scope of the label box so

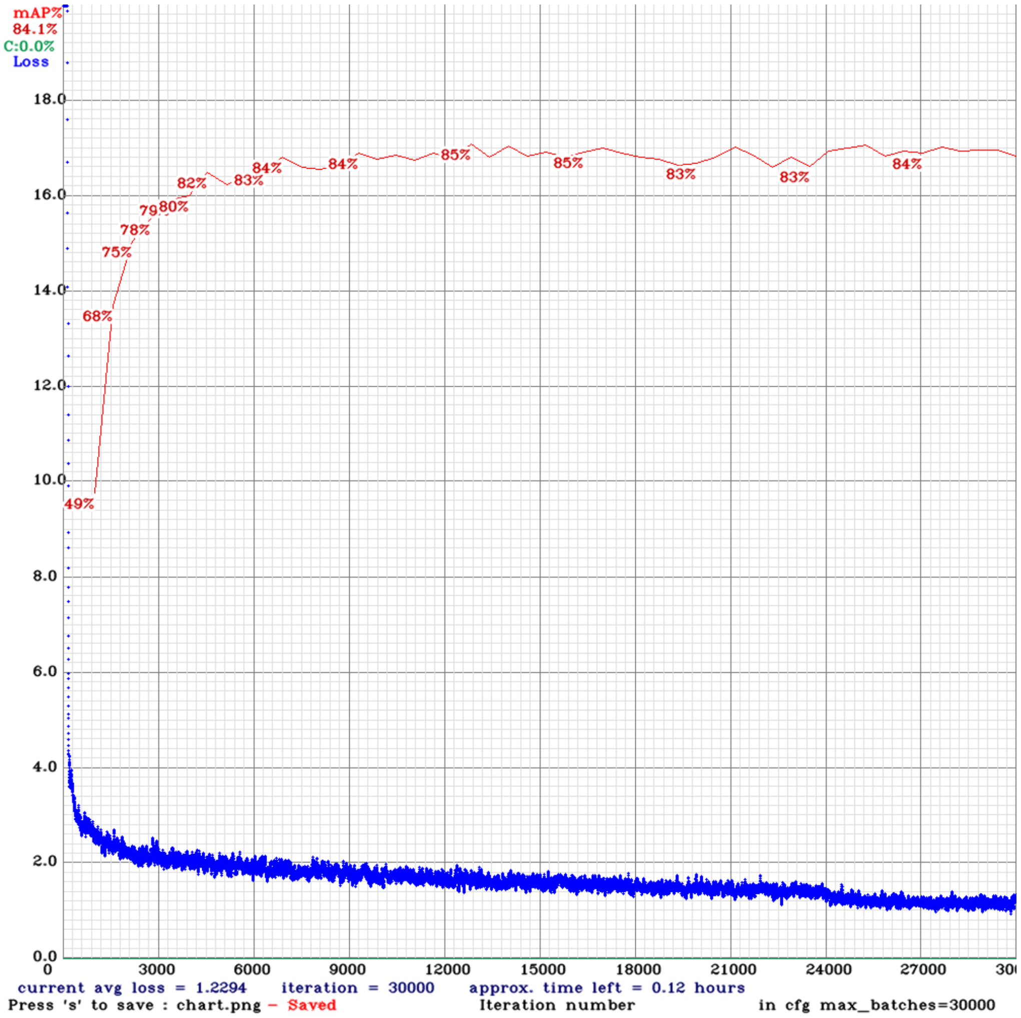


Figure 4.6: The training process of the partial labeling method.

that the vas deferens tissue in the image could be included in the label box as much as possible. Although the proportion of non-vas deferens in the label box increases, when we use the same training image, validation image and the same training parameters to train and test the model again, the training process and results show that the training effect of the new labeling method is better than our previous labeling method. The example after adjusting the labeling method is shown in line 2 of Figure 4.2, and the data details of the training process are shown in Figure 4.6 and Figure 4.7. In the figures, red line shows internal validation precision, blue line shows the loss curve.

Through the observation of the model training process, it is not difficult to find that different annotation methods will significantly affect the training effect of the

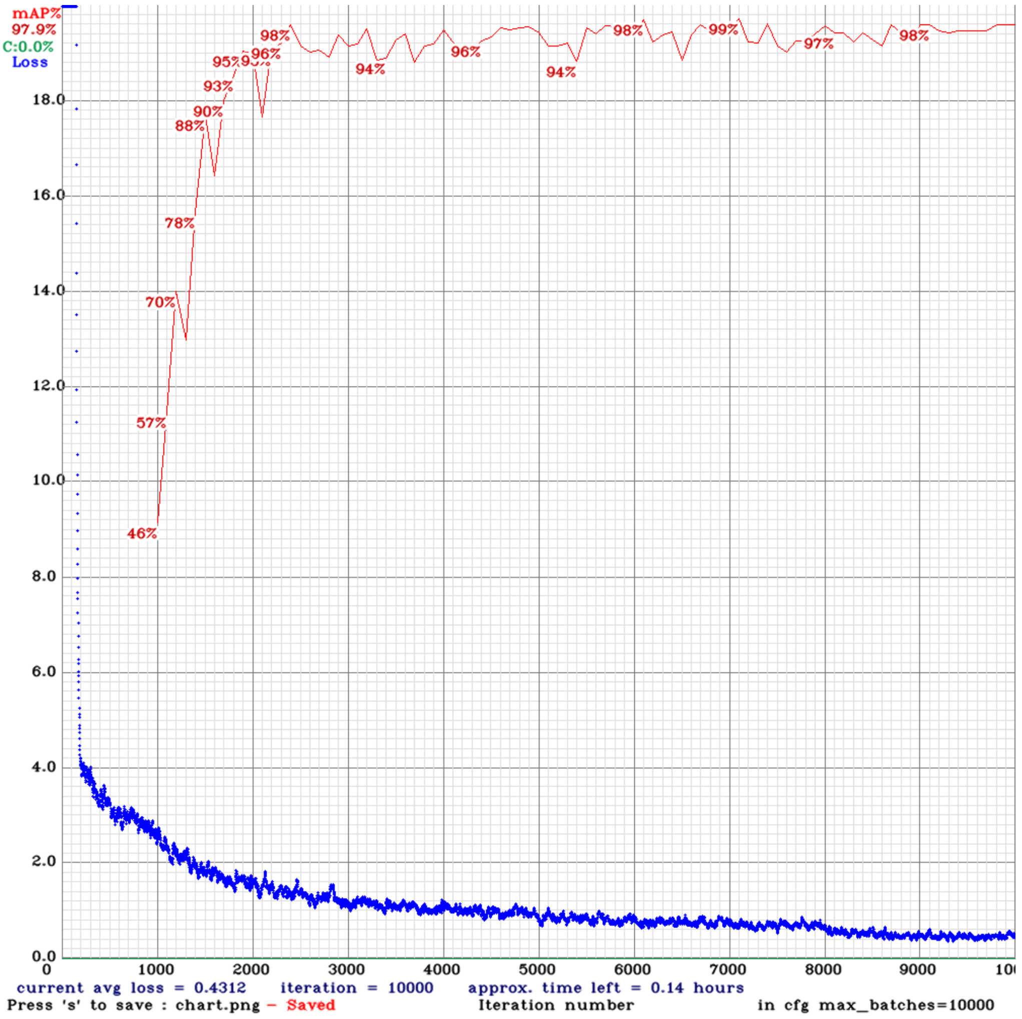


Figure 4.7: The training process of the new labeling method.

model. During surgery, although the target tissue may be partially occluded by other nontarget tissues, our suggestion is to label the target tissue as completely as possible.

4.4.3 Indicators IoU and AP

IoU is the ratio of the overlap area and the union area of the label box and the test box. In the field of medical image recognition, the higher the IoU, the higher the positioning accuracy, and the more it meets the needs of clinicians.

We tested the performance of the model under different IoU thresholds (0.7 to 0.2). The results showed that AP increased from 36.97% to 95.64%, and PPV increased from 58.82% to 99.82%.

We discussed with laparoscopic experts whether these labeled areas are enough to remind surgeons to identify the vas deferens. Experts believe that in laparoscopic surgery, the role of computer-aided surgery is to remind surgeons to readily discover target tissue and to pay more attention to this target area. When surgeons know the general location of the vas deferens, they will be more alert and cautious, so as not to damage the vas deferens and other target tissues during surgery.

By comparing the images in the test results, laparoscopic experts believed that when the IoU was greater than 0.3, the labeling result was acceptable. However, when the IoU dropped to 0.2, some labels were inaccurate, the proportion of target tissue in the label box was too low to correctly represent the vas deferens, and the center of the label box was not located over the vas deferens. These results also tell us that IoU is also an important indicator to evaluate the performance of the model, and higher IoU will better assist doctors to observe and discover target organs.

We found that when the IoU threshold was greater than 0.2, the sensitivity was lower than 90%, which indicated that the label box given by the model was not accurate enough and the learning of vas deferens was not enough. Although

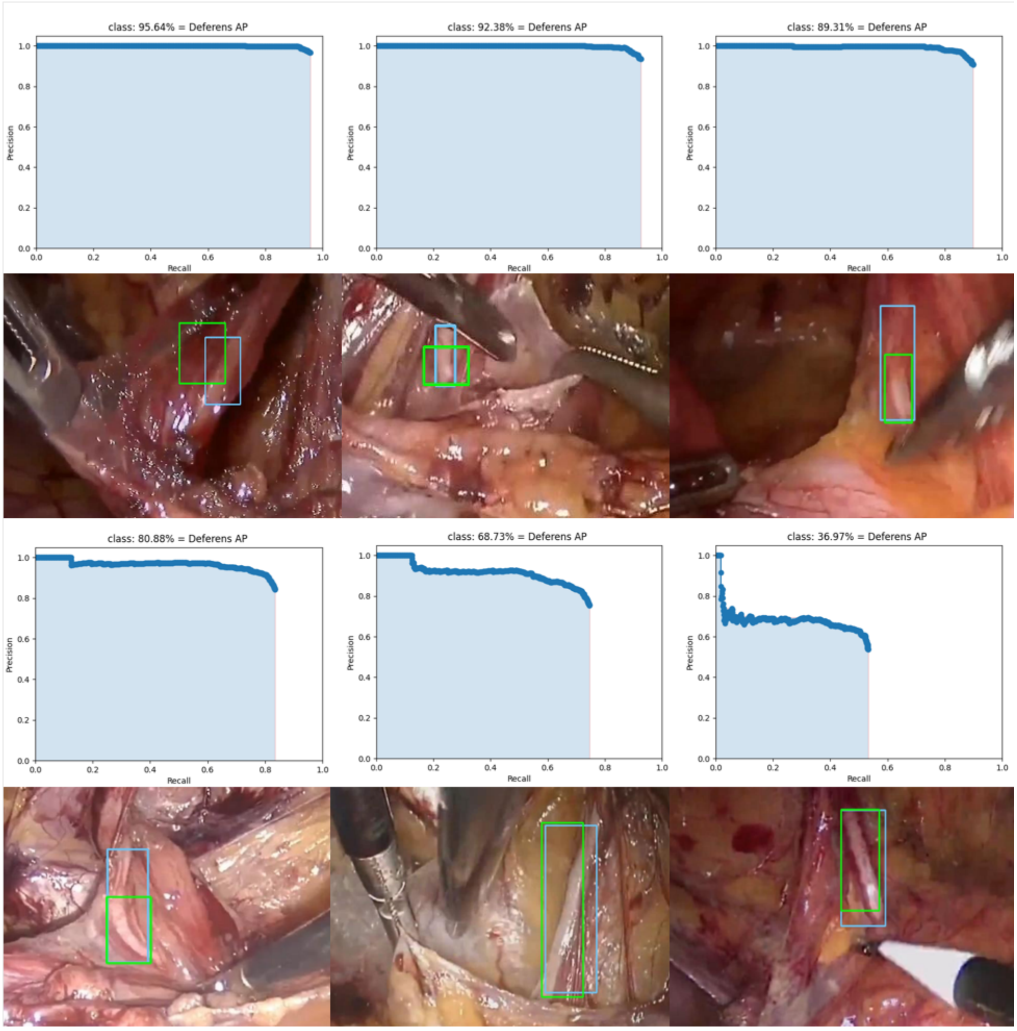


Figure 4.8: The AP of the model under different IoU values and the corresponding image examples. In the first row and the third row, from left to right, AP curves at IoU 0.2 to 0.4, and 0.5 to 0.7 are shown. In the second row and the fourth row, the corresponding images at different IoU values are shown. The blue rectangle boxes were labeled by laparoscopic experts in advance, and the green rectangle boxes were labeled by the model.

the model could recognize obvious features of the vas deferens, if the vas deferens' surface was partially obscured, the model could not accurately recognize and label the vas deferens. The details are shown in Figure 4.8.

Our research has some limitations, Firstly, we should further expand the number of patients and the number of vas deferens images as the training dataset, so that CNN model can fully learn the characteristics of vas deferens; Secondly, more indicators and test data should be used to evaluate the performance of the model, and we need to compare the performance with surgeons in different level, so as to make the evaluation of the model more objective. Thirdly, we only use the data of one hospital to train and verify the model, and the multi center data will more effectively prove the detection and generalization ability of the model. Finally, although Yolov4 is a good CNN model for target detection, we need to test more CNN models and compare their performance in order to select the best model to improve the detection performance.

In future studies, we plan to collect more data from more hospitals, compare the neural network with the indicators of identifying important tissues in laparoscopic inguinal hernia repair by surgeons with different levels of experience, and further test whether this technology can help young general surgeons optimize the learning curve and reduce the incidence of vas deferens injury complications.

4.5 Summary

As an effective object detection method, computer deep learning technology has been widely used in medical image recognition. [80, 102–104]

In this study, we used Yolov4 as the basic model to identify vas deferens images under TAPP and TEP. We used different confidence levels from 0.1 to 0.9 to calculate various evaluation indicators in image test dataset, picked the best confidence level for video test dataset, adjusted the IoU thresholds from 0.2 to 0.7 to understand the positioning accuracy and AP of the model and discussed with laparoscopic experts to select appropriate parameters to evaluate the performance of the model.

In image test dataset, the values of sensitivity, specificity, PPV, accuracy and F1 were 90.61%, 98.67%, 98.57%, 94.61% and 94.42% (confidence level 0.4), respectively. In video test dataset, the values of sensitivity, specificity, PPV, accuracy were 90.11%, 95.76%, 95.52%, 92.93% respectively. In IoU 0.3, the average precision (AP) was 92.38%.

We confirmed that we can use CNN model as an effective method to recognize vas deferens under TAPP and TEP. This will help laparoscopic surgeons, especially young ones, to better carry out clinical work, optimize the learning curve of laparoscopic surgery, improve surgical efficiency, and reduce surgical complications.

Chapter 5

Yolov4 based on Self-Attention Mechanism and Depthwise Separable Convolution

5.1 Introduction

The science of solving clinical problems by analyzing the images generated in clinical work is called medical image analysis. The aim is to extract important information effectively and improve the level of clinical diagnosis.

With the rapid development of Biomedical Engineering, medical image analysis has become one of the most popular research and development fields. Image classification using deep learning is the earliest and prosperous topic. Among them, CNN is the most widely used structure.

As we have studied in Chapter 3, eight popular CNN models have been used to train and test nerve and dura mater images under PTED, and good results have been obtained. However, the existing deep neural network model has high computational cost and high requirements for computer memory, which is difficult to be effectively implemented in devices with low memory resources or applications with real-time detection requirements.

In the field of object detection, detection speed and precision are very important indicators. Although Yolov4 has a good detection precision on our nerve and dura mater dataset, it still needs be improved on these two aspects.

Although all CNN models we used in Chapter 3 have good performance on our dataset, all these CNN models have a large number of parameters, and training these models is a time-consuming and laborious process. How to reduce computational complexity of the model and maintain the detection performance of the model is another challenge we need to try. Therefore, our proposal is to perform model compression and acceleration in existing deep networks without significantly reducing model performance.

Fortunately, in recent years, various CNN optimum operators have been developed. For reducing the parameter of CNN models, there are many efforts have been devoted to optimize/accelerate the inference speed of CNNs. These technologies attempt to improve the speed of the model in the training and prediction stages from the aspects of calculation optimization, system optimization and hardware optimization.

Calculation optimization technology is mainly to find a balance between model effect and efficiency, and reduce the amount of calculation of the model as much as possible while ensuring the model effect.

At present, the mainstream research mainly involves the following four technical schemes. The first one is model structure optimization. While the neural network model is developing in a wider and deeper direction, the model structure needs to be optimized to adapt to the current data and computational conditions. Most of the methods of model structure optimization are still based on human experience to design some ”light” computing components with similar effects to replace the ”heavy” computing components in the original model. The representative networks: VGG [96], NASNet [105], SqueezeNet [106], MobileNets [107], ShuffleNets [108] etc. The second one is model pruning. It can be divided into two categories: one is to cut the parameter matrix regularly (structured pruning), and the other is to cut the original dense parameter matrix into sparse parameter matrix (unstructured pruning). In 2014, Gong et al. [109] shown that model pruning and quantization are effective in reducing network complexity and solving over fitting problems. The third one is model quantification. It compresses the original network by reducing the number of bits required to represent each weight parameter, so as to accelerate the calculation. The fourth one is knowledge distillation. It is essentially similar to transfer learning, except for the purpose of model compression. It takes the knowledge learned from the large model as the prior knowledge, then transfers the prior knowledge to the small-scale neural network, and deploys the small-scale neural network in practical application. Ba et al. [110] adopted the idea of knowledge distillation (KD) to compress the deep and wide network into a shallow network, in which the compressed model imitates the learning function of a complex model.

In neural networks, we know that the convolution layer obtains the output features through the linear combination of convolution kernel and original features. Since the convolution kernel is usually local, we usually stack convolution layers in order to increase the receptive field. In fact, this method is not efficient. At the same time, many tasks of computer vision affect the final performance due to the lack of semantic information.

Self-attention (S-A) mechanism can capture global information to obtain larger receptive field and contextual information [111]. Recent studies have shown that S-A mechanism may be a feasible alternative method to build image recognition model [112, 113]. S-A mechanism has been adopted in natural language processing. As the basis of powerful architecture, it replaces recurrent and convolution models in various tasks [114–117]. The development of effective S-A architecture in computer vision provides a promising prospect for building models with different and possibly complementary characteristics from convolutional networks [118].

In this study, we proposed two approaches to reconstruct the CNN model based on Yolov4. One approach is to use depthwise separable convolution (DSC) [119]

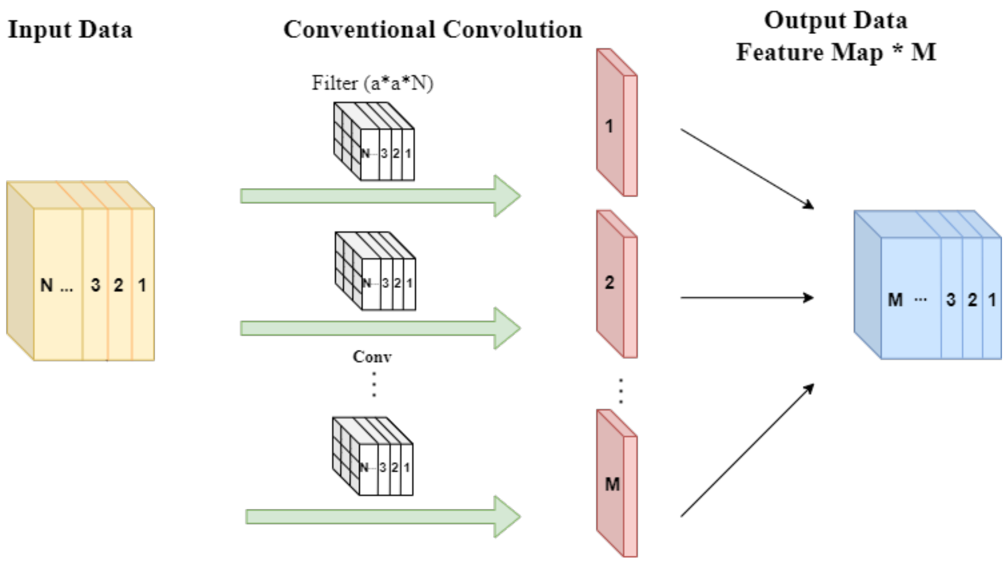


Figure 5.1: Process of conventional convolution.

to replace conventional convolution in order to reduce the parameter and the amount of calculation. Another approach is to use the S-A mechanism in the feature extraction process to model the global dependence of the input data.

The research shows that DSC is an effective method to reduce the amount of calculation. Compared with the conventional convolution, DSC can effectively reduce the calculation parameters, but the effect of target feature extraction is very close at the same time [119, 120]. The process of conventional convolution complete convolution calculation in one step (Figure 5.1), but the process of DSC divide the process of calculation into two steps (Figure 5.2). Step 1 is depthwise convolution (DW), to use convolution kernel to extract features. One convolution kernel is responsible for one channel, and one channel is convoluted by only one convolution kernel. Step 2 is pointwise convolution (PW), to use 1*1*M kernel (M is the number of channels of the original input layer) for conventional convolution operation. The convolution operation in this step will weight the map in the previous step in the depth direction to generate a new feature map. Therefore, although the output feature maps are similar, the number of parameters can be significantly reduced.

Multi-Head Attention performs multiple attention functions called Scaled Dot-Product Attention with keys, values and queries pairs in parallel. In the linear

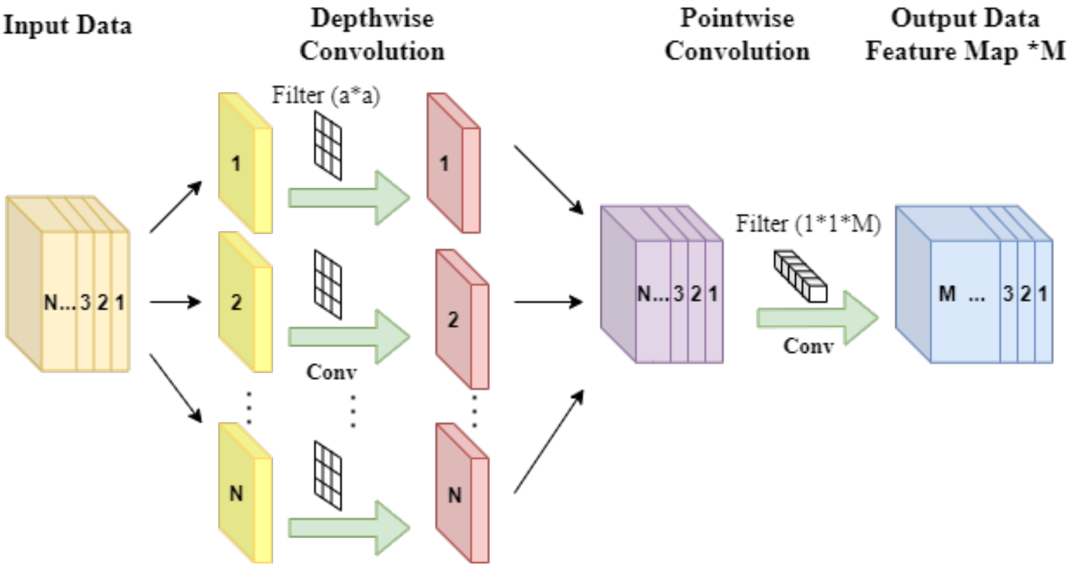


Figure 5.2: Process of depthwise separable convolution.

projection of these pairs, the process is performed with h times namely the number of heads, the h pairs are then transferred to Scaled Dot-Product Attention, which is described in Figure 5.3 [114]. The outputs of all heads are concatenated together as the final values. The Multi-Head mechanism allows the model to jointly attend to information from different representation sub-spaces at different positions.

As shown in Figure 5.4, in our case, the S-A process can be divided into three steps. Firstly, the input is transferred to three different convolutional operations for calculating Keys, Queries, and Values. Secondly, we compute the dot products of the queries and the keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the attention weights. Thirdly, the attention vectors can be calculated through the multiplication of the attention weights and values [114, 121]. The whole process can be formulated as:

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

As we studied before, CNN can as an effective approach to assistant surgeons recognize the important tissues under endoscopic surgery, including nerve and dura mater images under PTED, vas deferens images under laparoscopy. We compared 8 popular CNN models, and the Yolov4 has the best performance on nerve and

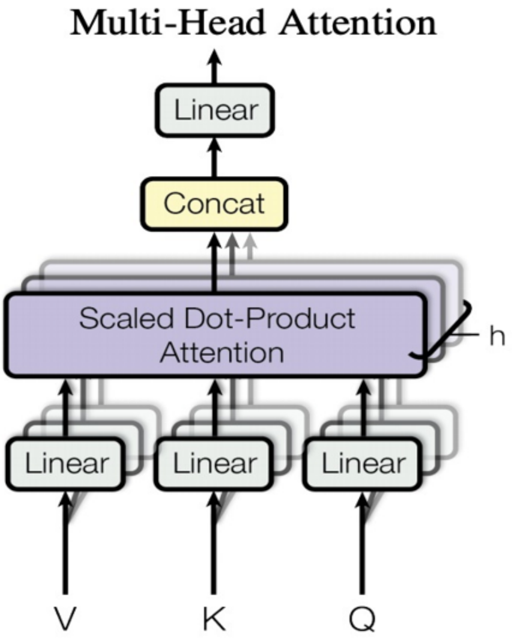


Figure 5.3: Illustration of MHSA.

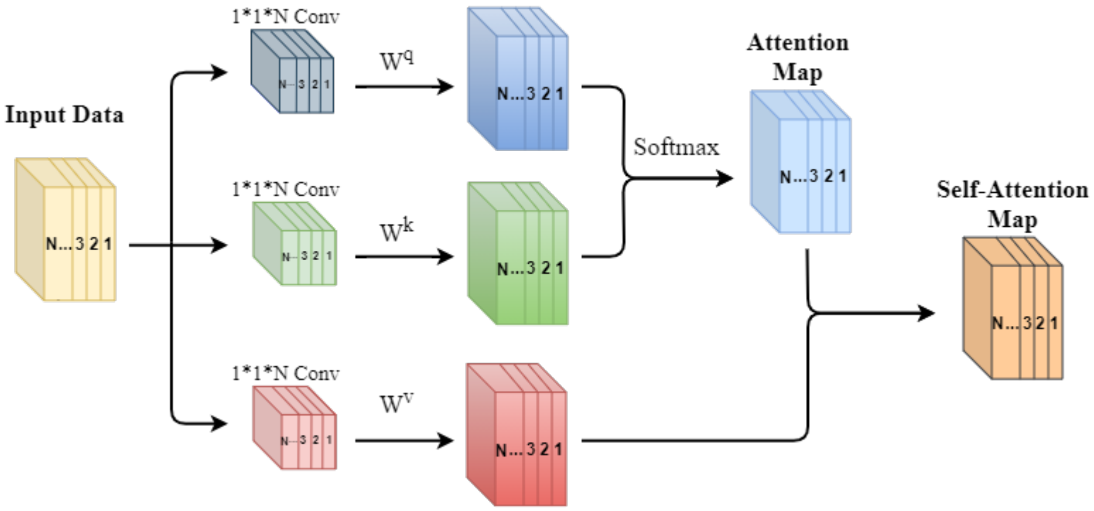


Figure 5.4: Illustration of self-attention mechanism.

dura mater images recognition under PTED. The sensitivity is 92.69%, and the specificity is 96.14%. It is a good performance that closely to the spinal endoscopy experts whose sensitivities are higher than 96%, and the specificities are higher than 98%.

Yolov4 as a classic one-stage CNN model, it has a good performance on COCO dataset, VOC dataset and other public datasets, but for our special medical image dataset, the object features and types maybe have some different with daily life images. So, we want to reconstruct the CNN model based on Yolov4 to reduce the model parameters and at the same time, we can maintain or improve the model's performance.

We roughly divided the structure of Yolov4 into three mainly parts: part 1 is the backbone of the model, the function of this part is to use conventional convolution and down sampling to impress and extract the preliminary features of the input data, then, it will extract 3 preliminary effective feature layers as the input features for next part; part 2 enhances feature extraction. It will merge the 3 preliminary effective layers and extract more effective features. In this study, we did not change the structure in this part; part 3 is a prediction network, it will output the classification and prediction results according to the part 2.

The innovations of this study are as follows: (1) DSC is used to replace the conventional convolution in Yolov4 backbone to reduce parameters, improve detection speed and test the performance of the model; (2) Add S-A mechanism to the three preliminary feature extraction layers in the Yolov4 backbone to improve the receptive field in the feature extraction process, so as to enhance the overall feature extraction of the data. (3) DSC combined with S-A mechanism was used to extract the preliminary features of the training data. The performance of the model in nerve and dura mater image recognition in PTED was evaluated from the aspects of sensitivity, specificity, model parameters, detection speed and AUC.

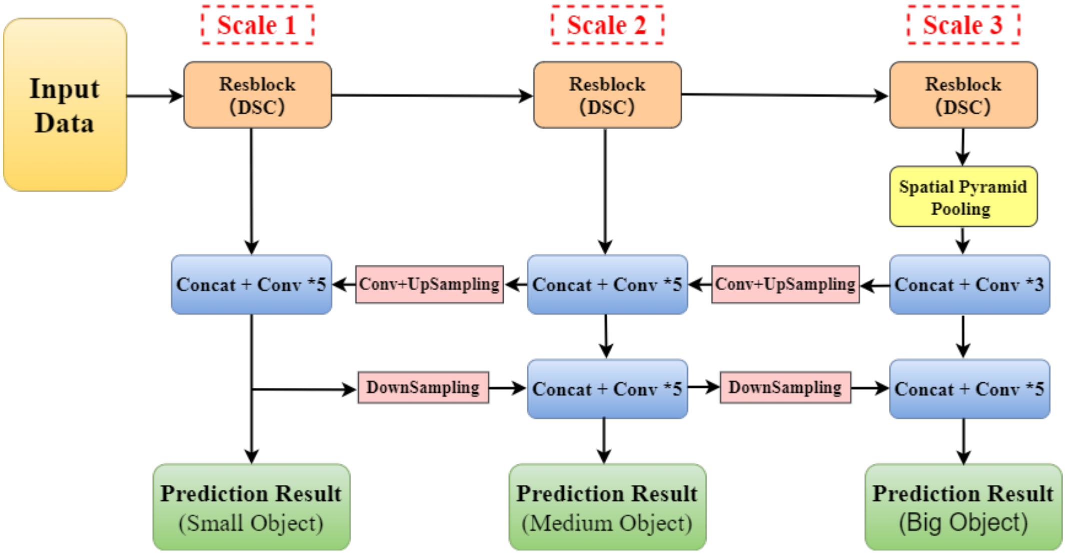


Figure 5.5: Structure of model 1.

5.2 Materials and Methods

5.2.1 CNN Models

All models are based on Yolov4, and we mainly adjust its backbone structure (CSPDarknet53). Firstly, three preliminary features are extracted and output through the backbone, and then the feature extraction is enhanced by conventional convolution, down sampling, up sampling and so on. Finally, according to different scales, the prediction results are outputted through three feature layers.

Model 1: Use DSC to replace the conventional convolutions in the backbone of Yolov4. The model structure is shown in Figure 5.5.

Model 2: DSC is used to replace the conventional convolution in the backbone and an S-A mechanism was added for preliminary feature extraction. The model structure is shown in Figure 5.6.

Model 3: S-A mechanism is added to the output of the three preliminary feature extraction layers of Yolov4. The model structure is shown in Figure 5.7.

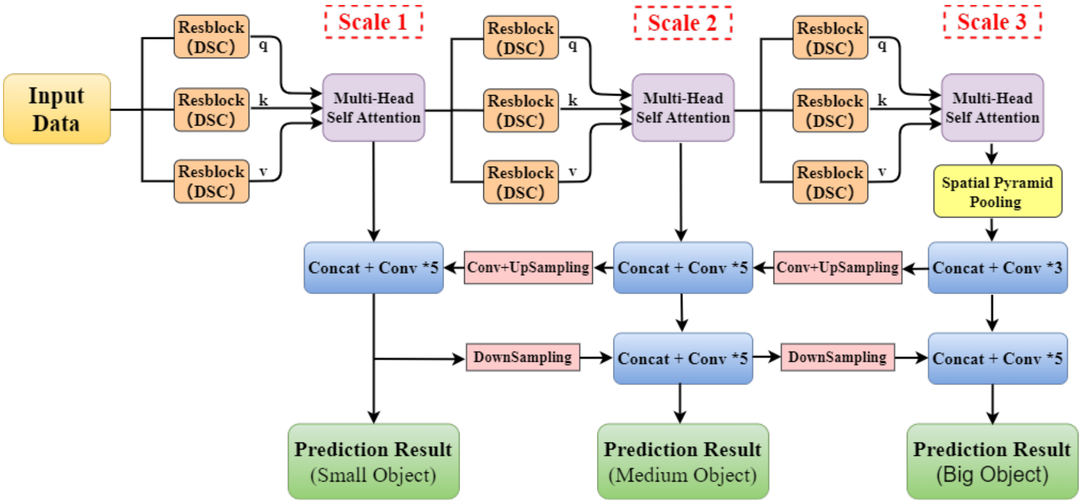


Figure 5.6: Structure of model 2.

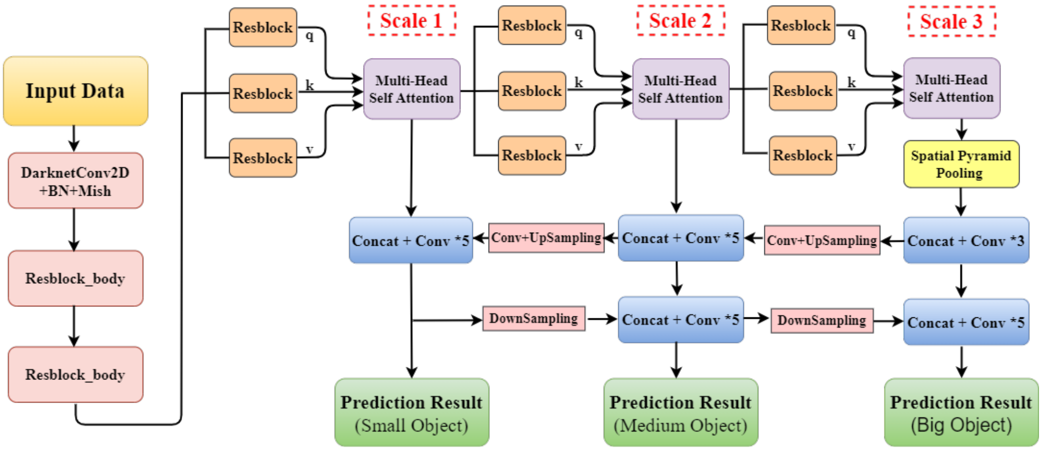


Figure 5.7: Structure of model 3.

5.2.2 Datasets

We used the dataset same as in Chapter 3, the details of information is shown in Table 3.1. There are 81 patients' 16,200 images including nerve and dura mater image in the training dataset, and then use these images as input data and be divided into training data and internal validation data randomly with a ratio 9:1. After training the model 200 epochs and observed the loss curves were no longer to reduce, then stop training. Finally, 21 patients' 8,400 images (4,200 with nerve and dura mater, 4,200 without nerve and dura mater) were used to test these models.

5.2.3 Computer Configuration and Training Parameters

The configuration was described in Chapter 3.2.3. For all CNN models, we use the same dataset to train and test these models, the training parameters are as follow: Input size, 608*608; Total epoch, 200; Freeze epoch, 50; Initial learning rate, 0.001; Unfreeze epoch 150; Learning rate, 0.0001.

5.2.4 Evaluation Indicators

In order to evaluate and compare the performance of these models, we still use common indicators, such as Sen, Spe, Acc, ROC curve, AUC value, FPS and the number of model parameters. The formulas are shown in Chapter 3.

5.2.5 Validation

Three spinal endoscopy experts reviewed all the detection results detected by the four models, checked the images and labels one by one, discussed the questionable results, and reached a consensus on the final results.

Table 5.1: Performance of the four models on the test dataset.

CNN Model	Sen(%)	Spe(%)	Acc(%)	PRM (10^6)	FPS	AUC
1	83.64	95.52	89.58	12.27	189.33	0.898
2	96.21	90.79	93.50	18.68	48.84	0.960
3	95.52	92.26	93.89	195.75	16.40	0.951
4	92.69	96.14	94.42	63.94	28.12	0.952

Note: PRM: number of parameters.

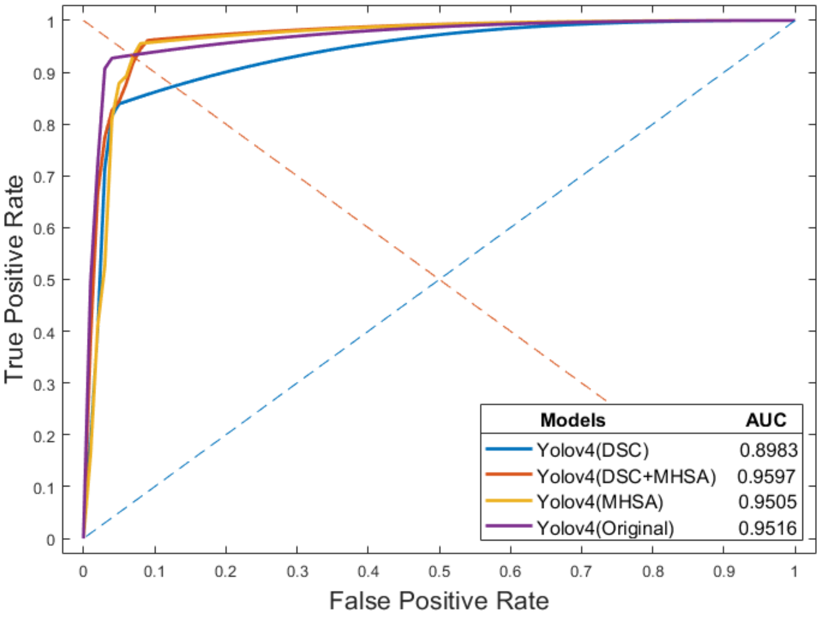


Figure 5.8: ROC curves and AUC values of four models on the test dataset.

5.3 Results

After the three spinal endoscopy experts checked and made a consent with the detection results output by CNN models, we confirmed the final results, and calculate the related evaluation indicators according the formulas shown in Chapter 3. The final results are shown in Table 5.1. The ROC curve and AUC value shown in figure 5.7, the radar plot of evaluation indicators of the models in test dataset respectively. We plot a more intuitive linear graph in Figure 5.8 to compare the performance of each model in different evaluation indicators.

These models can show good detection ability in the dataset, and the AUC value is between 0.8983-0.9597. The best AUC value 0.9597 was generated by

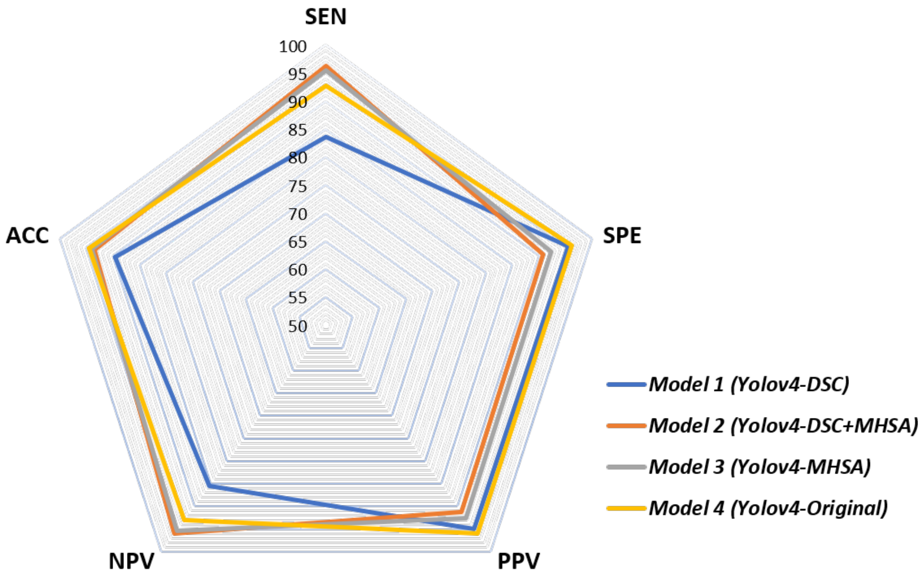
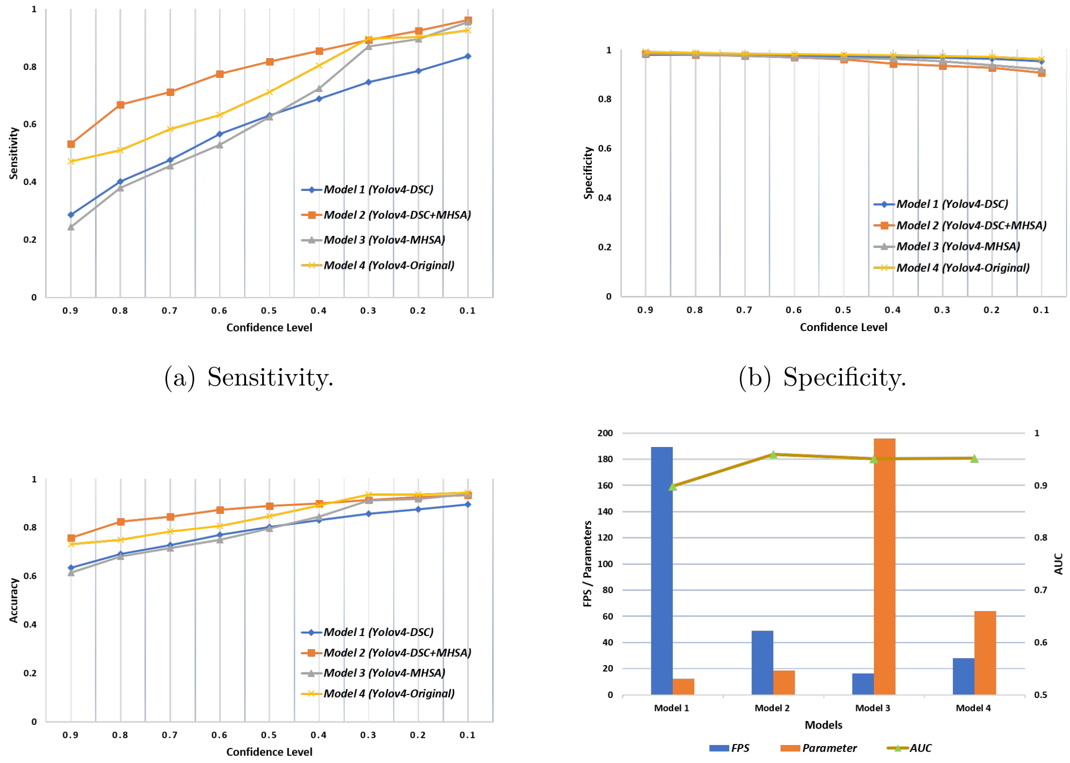


Figure 5.9: Overall performance of four models on the test dataset.



(a) Sensitivity.

(b) Specificity.

(c) Accuracy.

(d) FPS, parameters, and AUC values.

Figure 5.10: Comparison of the indicators of the four models.

model 2. The best sensitivity 96.21% was also generated by model 2, and the best specificity was 96.14%, which was generated by model 4.

We used Chi-square test to analyze the performance of model 2 and model 4. The p values of model 2 and model 4 were: sensitivity, $p < 0.001$; specificity, $p < 0.001$; accuracy $p = 0.003$ (p <0.05), which indicates that there are statistical differences in sensitivity, specificity and accuracy between model 2 and model 4. For sensitivity, model 2 is better than model 4, but for specificity and accuracy, model 4 is better than model 2.

In the comparison of model parameters, model 1 is the lightest model with parameter number of 12.27 $\times 10^6$. At the same time, the detection speed of this model is also the fastest. The original PTED video is 25 fps. In addition to model 3, model 1, model 2 and model 4 can meet the requirements of real-time detection.

5.4 Discussion

In this research, the DSC and S-A mechanism were used to reconstruct the CNN model Yolov4.

1. In order to reduce the size of the model and the calculation complexity, we used DSC to replace the conventional convolution in the backbone of Yolov4. Results shown that the number of parameters of the model were reduced from 63.94 $\times 10^6$ to 12.27 $\times 10^6$, the detection speed increased from 28.12 fps to 189.3 fps. But the detection accuracy was decreased from 94.42% to 89.58%, and the AUC value decreased from 0.95 to 0.90. These results show that when DSC is used to replace some conventional convolution, the number of parameters of the model is reduced to only one quarter of Yolov4, and the performance of the model on our data set is also acceptable.

2. While using DSC to replace some conventional convolution, we add S-A mechanism to the three outputs of the backbone. The results show that the number of parameters of the model is much fewer than that of the original Yolov4

model, only one third of that of Yolov 4 model, and the detection speed can reach 48.84 fps, which can meet the needs of most medical endoscopes or surgical videos. For detection accuracy, the model 2 is lower than original Yolov4. According to the results showed in Table 1 and Figure 8a, for sensitivity, although the best sensitivity of model 2, 3, and 4 are close, but when we observed the sensitivity curve, even the confidence level was higher than 0.9, the sensitivity of the model 2 is still higher than 50%. Although the specificity of model 2 is lower than the model 4, but we find that the AUC value 0.96 is also higher than model 4. It is also indicated that, for the comprehensive performance, the capability of model 2 is higher than model 4 （original Yolov4）.

3. In the process of building model 3, we added S-A mechanism to the backbone of Yolov4 directly. Because each point must capture the global context information, the self-attention mechanism module will have great computational complexity and memory capacity [114]. We compared the detection performance of model 3 and model 4, although the best performance was similar in these two models, the AUC value was also similar to each other. The number of parameters rose from 64 $\times 10^6$ to 3 times 196 $\times 10^6$, with the detection speed reduced from 28.12 to 16.40 fps.

The limitations of this study are given below. First, performance of the proposed CNN models was not compared with spinal endoscopy doctors. We will further to compare the performance of these CNN models to spinal endoscopy doctors. Second, the resource of the dataset is relatively single, and the sample size is not large enough. So, future studied should plan to further expand the data source and sample size. Third, the detection accuracy of the reconstructed model didn't surpass the original Yolov4. We should do more study on the model reconstruct, and acquire the better performance.

5.5 Summary

In summary, CNN is an effective method to recognize nerve and dura mater images under PTED. For model building and reconstruction, there are many methods and approachs to do.

In this study, the use of DSC combined with S-A mechanism can significantly reduce the model parameters and improve the detection speed of the model. At the same time, the overall performance of the reconstructed model in identifying nerve and dura mater in PTED is similar to that of Yolov4. This shows that this lightweight CNN model also has a good effect on medical image recognition.

As we reviewed before, there are many ways to achieve network lightweight. For future work, we can try to use other methods to simplify the scale of CNN model, and we can also use these technologies to build personalized networks for different medical datasets. This is also a popular method called precision medicine at this stage. It will not only bring efficient assistance to the medical workers, but also make patients more assured and safer in the process of medical treatment.

Chapter 6

Conclusion and Future Works

By reviewing the application of deep learning technology, especially CNN technology in various medical activities in recent years, we find that the technology has achieved good performance in various fields. Its application covers a wide range of issues from disease screening, disease monitoring, lesion detection to personalized treatment recommendations. It opens up an unprecedented new field for medical data processing and analysis.

In the background of the popularity of medical big data, a variety of data sources, such as radiography (X-ray, CT and MRI scanning), endoscopy / surgery video, ECG / EEG signal, pathological imaging and genome sequence, bring a lot of data to doctors.

However, in medical image analysis, useful information is not only contained in the image itself. Doctors usually make better decisions by using a large amount of data about the patient's medical history, age, occupation and so on. We also need to use more tools to analyze these data in depth, so as to extract more useful information and promote the development of related disease research and medicine.

This dissertation concludes with a summary of the main contributions and a discussion of future research. We create an overview of the results of each contribution. The future work will continue to take computer technology as the core to further understand and meet the needs of clinicians, and design more conducive to clinical practical applications.

6.1 Contributions and Conclusion

In Chapter 2, we use a CNN model based on Yolov3 to train and test the recognition ability of nerve and dura mater images under PTED. We compare the performance with different level surgeons, and as another evaluate indicator to estimate the performance of the CNN model. We combined the CNN to each surgeon, use statistical method to evaluate the assistance effect of the CNN model.

This is the first time we used CNN model to identify the nerve and dura mater under the PTED, the test results indicated that CNN could recognize the tissues image during the PTED, and this is an efficient method for nerve and dura mater recognition and can assist general surgeons during the PTED.

In Chapter 3, in order to further validate if the CNN model can identify the nerve and dura mater effectively under the PTED, we tested eight different popular CNN models (include one-stage and two-stage) to complete the nerve and dura mater detection. We amplified the research dataset and trained the eight different CNN models using the similar parameters, respectively.

We compared the results with different CNN models, and evaluated the performance of the CNN models with the same indicators. According to the detection results, we find that all of these CNN models can recognize nerve and dura mater effectively, the best AUC is 0.972, produced by SSD, the best precision is 94.42%, produced by Yolov4, the fastest model is Yolov4-Tiny, the FPS can reach 196.15. These results indicated that the CNN model is an effective tool for nerve and dura mater detection, it can assist surgeons to recognize the nerve and dura mater accurately and immediately.

In Chapters 2 and 3, we aimed to use CNN models to recognize the nerve and dura mater images under the PTED. The goal is to help reduce the workload of spinal endoscope experts and improve the performance of general surgeons. Results shown that CNN model has higher sensitivity and specificity than general surgeons. When general surgeons are combined with CADs, the performance of general surgeons is significantly improved. It also shown that CNN model can effectively identify nerves and dura under PTED.

In Chapter 4, we used CNN model based on Yolov4 to identify the vas deferens images under laparoscopic hernial repair surgery. We used different labeling methods to label the training data to provide more effective training features for the training model, so as to improve the training effect and test accuracy of the model. We also tested the real-time detection ability of the model using video clips (30 fps). The test results were TPR 90.61%, TNR 98.67%, Acc 94.61%. The results also shown that this CNN model can recognize the vas deferens in images and videos.

We confirmed that CNNs can recognize and label vas deferens images efficiently in TAPP and TEP. This will assist laparoscopic surgeons, especially young ones, to better carry out clinical work, optimize the learning curve of laparoscopic surgery, improve surgical efficiency, and reduce surgical complications.

In Chapter 5, Yolov4 was reconstructed by DSC and S-A mechanism. When we use DSC to replace the conventional convolution in the backbone of Yolov4 and add S-A mechanism in the three preliminary feature layers respectively, the model parameters are reduced to 18.68 $\times 10^6$, the detection speed is increased to 48.84 FPS, the AUC value of the model is 0.96, and the sensitivity is 96.21%. The performance of other indicators is similar to Yolov4.

It also shows that for a specific medical image field, using the existing technology to design a personalized lightweight network for specific medical image detection can also receive good performance.

6.2 Future Trends and Works

With the rapid development of deep learning technology, especially the method based on CNN, the CAD system based on CNN has a wider application prospect in clinical practice. In the foreseeable future, these technologies are not expected to replace doctors, but may promote daily workflow, optimize doctors' learning curve for new technologies, improve the accuracy of detection and diagnosis, make the application of minimally invasive and endoscopic technology more and more popular, reduce invasive operation, reduce medical operation risk and improve patient satisfaction.

Deep learning techniques extract information from big data and produce an output that can be used for personalized treatment, which promotes the development of precision medicine. Unlike the conventional medical treatment, in precision medicine, the examination will go deep into the smallest molecular and genomics information, and medical staff will make the diagnostic decision according to the subtle differences among patients.

With the continuous upgrading of powerful computing equipment, object detection technology based on deep learning has developed rapidly. In order to deploy in more accurate applications, the demand for high-precision real-time systems becomes more and more urgent.

Although the application of CNN in the medical field has achieved great results in recent years, there are still many unknown areas that are worthy of further exploration and development.

For the future research, I also have some plans to do.

The first one is to combining one-stage and two-stage detectors. The two-stage detectors have s densely training process to obtain as many as reference boxes, it is time consuming and inefficient. To resolve this issue, we need to eliminate the redundant detection boxes and maintain the high accuracy. The one-stage detectors achieve fast processing speed and have been applied in real-time object detection. Although the detection speed is fast, the lower accuracy is still a bottleneck for high precision requirement. How to combine the advantages of both two-stage and one-stage detectors remain a big challenge.

The second one is to deal with focus on low-quality images. In the process of target detection, we find that the accuracy of the model is low in the detection of blurred and defocused videos, small targets and so on. In such condition, it is difficult to achieve good performance in practical clinical application. How to improve the detection accuracy in these cases is also my research direction.

The third one is to build a lightweight object detection model. Starting from the definition of lightweight network structure, we can divide lightweight networks into two categories: lightweight network structure design and model compression. At this stage, in addition to some model compression methods, researchers pay more attention on the manual design of lightweight and efficient CNN structure, which can effectively maintain the accuracy of the model and greatly reduce the parameters. How to build an optimal network structure for medical target detection is also my goal.

In short, based on the existing computer technology and using the professional knowledge we have learned, the application of artificial intelligence in the medical field is becoming wider and safer.

References

[1] G. B. Andersson, “Epidemiological features of chronic low-back pain,” *The lancet*, vol. 354, no. 9178, pp. 581–585, 1999.

[2] D. Hoy, C. Bain, G. Williams, L. March, P. Brooks, F. Blyth, A. Woolf, T. Vos, and R. Buchbinder, “A systematic review of the global prevalence of low back pain,” *Arthritis & Rheumatism*, vol. 64, no. 6, pp. 2028–2037, 2012.

[3] G. K. Lutz, M. Butzlaff, and U. Schultz-Venrath, “Looking back on back pain: trial and error of diagnoses in the 20th century,” *Spine*, vol. 28, no. 16, pp. 1899–1905, 2003.

[4] H.-T. Hsu, S.-J. Chang, S. S. Yang, and C. L. Chai, “Learning curve of full-endoscopic lumbar discectomy,” *European Spine Journal*, vol. 22, no. 4, pp. 727–733, 2013.

[5] K. S. Cahill, A. D. Levi, M. D. Cummock, W. Liao, and M. Y. Wang, “A comparison of acute hospital charges after tubular versus open microdiskectomy,” *World neurosurgery*, vol. 80, no. 1-2, pp. 208–212, 2013.

[6] E. G. Manusov, “Surgical treatment of low back pain,” *Primary Care: Clinics in Office Practice*, vol. 39, no. 3, pp. 525–531, 2012.

[7] A. T. Yeung and P. M. Tsou, “Posterolateral endoscopic excision for lumbar disc herniation: surgical technique, outcome, and complications in 307 consecutive cases,” *Spine*, vol. 27, no. 7, pp. 722–731, 2002.

[8] Y. Ahn, S.-H. Lee, W.-M. Park, H.-Y. Lee, S.-W. Shin, and H.-Y. Kang, "Percutaneous endoscopic lumbar discectomy for recurrent disc herniation: surgical technique, outcome, and prognostic factors of 43 consecutive cases," *Spine*, vol. 29, no. 16, pp. E326–E332, 2004.

[9] D. Y. Lee, C. S. Shim, Y. Ahn, Y.-G. Choi, H. J. Kim, and S.-H. Lee, "Comparison of percutaneous endoscopic lumbar discectomy and open lumbar microdiscectomy for recurrent disc herniation," *Journal of Korean Neurosurgical Society*, vol. 46, no. 6, p. 515, 2009.

[10] S. Ruetten, M. Komp, H. Merk, and G. Godolias, "Recurrent lumbar disc herniation after conventional discectomy: a prospective, randomized study comparing full-endoscopic interlaminar and transforaminal versus microsurgical revision," *Clinical Spine Surgery*, vol. 22, no. 2, pp. 122–129, 2009.

[11] J. A. Gibson, J. G. Cowie, and M. Iprenburg, "Transforaminal endoscopic spinal surgery: the future 'gold standard' for discectomy?–a review," *the surgeon*, vol. 10, no. 5, pp. 290–296, 2012.

[12] T. Bandoh, N. Shiraishi, Y. Yamashita, T. Terachi, M. Hashizume, S. Akira, T. Morikawa, Y. Kitagawa, K. Yanaga, S. Endo *et al.*, "Endoscopic surgery in japan: The 12th national survey (2012–2013) by the japan society for endoscopic surgery," *Asian journal of endoscopic surgery*, vol. 10, no. 4, pp. 345–353, 2017.

[13] H. Shiroshita, M. Inomata, T. Bandoh, H. Uchida, S. Akira, M. Hashizume, S. Yamaguchi, S. Eguchi, N. Wada, S. Takiguchi *et al.*, "Endoscopic surgery in japan: The 13th national survey (2014-2015) by the japan society for endoscopic surgery," *Asian journal of endoscopic surgery*, vol. 12, no. 1, pp. 7–18, 2019.

[14] T. Hoogland, K. van den Brekel-Dijkstra, M. Schubert, and B. Miklitz, "Endoscopic transforaminal discectomy for recurrent lumbar disc herniation:

a prospective, cohort evaluation of 262 consecutive cases," *Spine*, vol. 33, no. 9, pp. 973–978, 2008.

[15] W. Hua, Y. Zhang, X. Wu, Y. Gao, S. Li, K. Wang, S. Yang, and C. Yang, "Full-endoscopic visualized foraminoplasty and discectomy under general anesthesia in the treatment of l4-l5 and l5-s1 disc herniation," *Spine*, vol. 44, no. 16, pp. E984–E991, 2019.

[16] R. J. Fitzgibbons Jr and R. A. Forse, "Groin hernias in adults," *New England Journal of Medicine*, vol. 372, no. 8, pp. 756–763, 2015.

[17] A. Kingsnorth, M. Gingell-Littlejohn, S. Nienhuijs, S. Schüle, P. Appel, P. Ziprin, A. Eklund, M. Miserez, and S. Smeds, "Randomized controlled multicenter international clinical trial of self-gripping parietex™ progrip™ polyester mesh versus lightweight polypropylene mesh in open inguinal hernia repair: interim results at 3 months," *Hernia*, vol. 16, no. 3, pp. 287–294, 2012.

[18] P. Kapur, M. G. Caty, and P. L. Glick, "Pediatric hernias and hydroceles," *Pediatric Clinics of North America*, vol. 45, no. 4, pp. 773–789, 1998.

[19] T. Vos, C. Allen, M. Arora, R. M. Barber, Z. A. Bhutta, A. Brown, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015," *The lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016.

[20] N. Stylopoulos, G. S. Gazelle, and D. Rattner, "A cost–utility analysis of treatment options for inguinal hernia in 1,513,008 adult patients," *Surgical Endoscopy And Other Interventional Techniques*, vol. 17, no. 2, pp. 180–189, 2003.

[21] I. Abubakar, T. Tillmann, and A. Banerjee, "Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death,

1990-2013: a systematic analysis for the global burden of disease study 2013," *Lancet*, vol. 385, no. 9963, pp. 117–171, 2015.

[22] H. Wang, M. Naghavi, C. Allen, R. M. Barber, Z. A. Bhutta, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen, M. M. Coates *et al.*, "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015," *The lancet*, vol. 388, no. 10053, pp. 1459–1544, 2016.

[23] E. Bassini *et al.*, "Nuovo metodo per la cura radicale dell'ernia inguinale," *Atti Congr Assoc Med Ital*, vol. 2, p. 179, 1887.

[24] R. Ger, K. Monroe, R. Duvivier, and A. Mishrick, "Management of indirect inguinal hernias by laparoscopic closure of the neck of the sac," *The American Journal of Surgery*, vol. 159, no. 4, pp. 370–373, 1990.

[25] C. C. Edwards, R. W. Bailey *et al.*, "Laparoscopic hernia repair: the learning curve," *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, vol. 10, no. 3, pp. 149–153, 2000.

[26] L. Neumayer, A. Giobbie-Hurder, O. Jonasson, R. Fitzgibbons Jr, D. Dunlop, J. Gibbs, D. Reda, and W. Henderson, "Open mesh versus laparoscopic mesh repair of inguinal hernia," *New England journal of medicine*, vol. 350, no. 18, pp. 1819–1827, 2004.

[27] M. Genitsaris, I. Goulimaris, and N. Sikas, "Laparoscopic repair of groin pain in athletes," *The American journal of sports medicine*, vol. 32, no. 5, pp. 1238–1242, 2004.

[28] E. Kuhry, R. Van Veen, H. Langeveld, E. Steyerberg, J. Jeekel, and H. Bonjer, "Open or endoscopic total extraperitoneal inguinal hernia repair? a systematic review," *Surgical endoscopy*, vol. 21, no. 2, pp. 161–166, 2007.

[29] N. J. Nilsson, *The quest for artificial intelligence.* Cambridge University Press, 2009.

[30] T. M. Mitchell *et al.*, "Machine learning," 1997.

[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[32] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets.* Springer, 1982, pp. 267–285.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[35] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[36] K. He and G. Gkioxari, "Dollá r p, girshick r. mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[38] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of*

the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[40] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[41] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision.* Springer, 2014, pp. 391–405.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[43] Y. Chen, C. Han, N. Wang, and Z. Zhang, "Revisiting feature alignment for one-stage object detection," *arXiv preprint arXiv:1908.01570*, 2019.

[44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[45] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9627–9636.

[46] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[47] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 840–849.

[48] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[49] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision.* Springer, 2016, pp. 21–37.

[50] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, “Region proposal by guided anchoring,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2965–2974.

[51] K. Kerlikowske, P. A. Carney, B. Geller, M. T. Mandelson, S. H. Taplin, K. Malvin, V. Ernster, N. Urban, G. Cutter, R. Rosenberg *et al.*, “Performance of screening mammography among women with and without a first-degree relative with breast cancer,” *Annals of internal medicine*, vol. 133, no. 11, pp. 855–863, 2000.

[52] K. Doi, “Computer-aided diagnosis in medical imaging: historical review, current status and future potential,” *Computerized medical imaging and graphics*, vol. 31, no. 4-5, pp. 198–211, 2007.

[53] F. Winsberg, M. Elkin, J. Macy Jr, V. Bordaz, and W. Weymouth, “Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis,” *Radiology*, vol. 89, no. 2, pp. 211–215, 1967.

[54] V. M. Rao, D. C. Levin, L. Parker, B. Cavanaugh, A. J. Frangos, and J. H. Sunshine, “How widely is computer-aided detection used in screening and diagnostic mammography?” *Journal of the American College of Radiology*, vol. 7, no. 10, pp. 802–805, 2010.

[55] J. D. Keen, J. M. Keen, and J. E. Keen, “Utilization of computer-aided detection for digital screening mammography in the united states, 2008 to

2016," *Journal of the American College of Radiology*, vol. 15, no. 1, pp. 44–48, 2018.

[56] F. Beyer, L. Zierott, E. Fallenberg, K. Juergens, J. Stoeckel, W. Heindel, and D. Wormanns, "Comparison of sensitivity and reading time for the use of computer-aided detection (cad) of pulmonary nodules at mdct as concurrent or second reader," *European radiology*, vol. 17, no. 11, pp. 2941–2947, 2007.

[57] O. Gautschi, G. Hildebrandt, and D. Cadosch, "Acute low back pain–assessment and management," *Praxis*, vol. 97, no. 2, pp. 58–68, 2008.

[58] P. Campbell, G. Wynne-Jones, S. Muller, and K. M. Dunn, "The influence of employment social support for risk and prognosis in nonspecific back pain: a systematic review and critical synthesis," *International archives of occupational and environmental health*, vol. 86, no. 2, pp. 119–137, 2013.

[59] P. Apostolides, R. Jacobowitz, and V. Sonntag, "Lumbar discectomy microdiscectomy:" the gold standard"," *Clinical neurosurgery*, vol. 43, pp. 228–238, 1996.

[60] M. Kim, S. Lee, H.-S. Kim, S. Park, S.-Y. Shim, and D.-J. Lim, "A comparison of percutaneous endoscopic lumbar discectomy and open lumbar microdiscectomy for lumbar disc herniation in the korean: a meta-analysis," *BioMed research international*, vol. 2018, 2018.

[61] Ahn and Yong, "Transforaminal percutaneous endoscopic lumbar discectomy: technical tips to prevent complications," *Expert Review of Medical Devices*, vol. 9, no. 4, pp. 361–366.

[62] S. J. Müller, B. W. Burkhardt, and J. M. Oertel, "Management of dural tears in endoscopic lumbar spinal surgery: A review of the literature," *World neurosurgery*, vol. 119, pp. 494–499, 2018.

[63] V. Puvanesarajah and H. Hassanzadeh, “The true cost of a dural tear: medical and economic ramifications of incidental durotomy during lumbar discectomy in elderly medicare beneficiaries,” *Spine*, vol. 42, no. 10, pp. 770–776, 2017.

[64] M. E. Murphy, J. S. Hakim, P. Kerezoudis, M. A. Alvi, D. S. Ubl, E. B. Habermann, and M. Bydon, “Micro vs. macrodiscectomy: does use of the microscope reduce complication rates?” *Clinical neurology and neurosurgery*, vol. 152, pp. 28–33, 2017.

[65] K. Kotil, “Closed drainage versus non-drainage for single-level lumbar disc surgery: relationship between epidural hematoma and fibrosis,” *Asian spine journal*, vol. 10, no. 6, p. 1072, 2016.

[66] Y. Nohara, H. Taneichi, K. Ueyama, N. Kawahara, K. Shiba, Y. Tokuhashil, T. Tani, S. Nakahara, and T. Iida, “Nationwide survey on complications of spine surgery in japan,” *journal of Orthopaedic Science*, vol. 9, no. 5, pp. 424–433, 2004.

[67] M. Liuke, S. Solovieva, A. Lamminen, K. Luoma, P. Leino-Arjas, R. Luukkonen, and H. Riihimäki, “Disc degeneration of the lumbar spine in relation to overweight,” *International journal of obesity*, vol. 29, no. 8, pp. 903–908, 2005.

[68] M. J. Perez-Cruet, R. G. Fessler, and N. I. Perin, “complications of minimally invasive spinal surgery,” *Neurosurgery*, vol. 51, no. suppl_2, pp. S2–26, 2002.

[69] L.-m. Rong, P.-g. Xie, D.-h. Shi, J.-w. Dong, L. Bin, F. Feng, and D.-z. Cai, “Spinal surgeons’ learning curve for lumbar microendoscopic discectomy: a prospective study of our first 50 and latest 10 cases,” *Chinese medical journal*, vol. 121, no. 21, pp. 2148–2151, 2008.

[70] J. A. Sclafani and C. W. Kim, “Complications associated with the initial learning curve of minimally invasive spine surgery: a systematic review,” *Clinical Orthopaedics and Related Research®*, vol. 472, no. 6, pp. 1711–1717, 2014.

[71] D.-J. Choi, C.-M. Choi, J.-T. Jung, S.-J. Lee, and Y.-S. Kim, “Learning curve associated with complications in biportal endoscopic spinal surgery: challenges and strategies,” *Asian spine journal*, vol. 10, no. 4, p. 624, 2016.

[72] S. Tenenbaum, H. Arzi, A. Herman, A. Friedlander, and I. Caspi, “Percutaneous posterolateral transforaminal endoscopic discectomy: Clinical outcome, complications, and learning curve evaluation,” *Surg Technol Int*, vol. XXI, pp. 278–283, 2012.

[73] P. Cui, Z. Guo, J. Xu, T. Li, Y. Shi, W. Chen, T. Shu, and J. Lei, “Tissue recognition in spinal endoscopic surgery using deep learning,” in *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*. IEEE, 2019, pp. 1–5.

[74] B. Kayalibay, G. Jensen, and P. van der Smagt, “Cnn-based segmentation of medical imaging data,” *arXiv preprint arXiv:1701.03056*, 2017.

[75] K. Suzuki, “Overview of deep learning in medical imaging,” *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.

[76] D. Feng, “An ensemble of fine-tuned convolutional neural networks for medical image classification,” *IEEE Journal of Biomedical Health Informatics*, vol. 21, no. 1, pp. 31–40.

[77] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[78] L. Nannia, S. Ghidoni, and S. Brahnam, “Ensemble of convolutional neural networks for bioimage classification,” *Applied Computing and Informatics*, 2020.

[79] P. Azimi, T. Yazdanian, E. C. Benzel, H. N. Aghaei, S. Azhari, S. Sadeghi, and A. Montazeri, "A review on the use of artificial intelligence in spinal diseases," *Asian Spine Journal*, 2020.

[80] H. Luo, G. Xu, C. Li, L. He, L. Luo, Z. Wang, B. Jing, Y. Deng, Y. Jin, Y. Li *et al.*, "Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study," *The Lancet Oncology*, vol. 20, no. 12, pp. 1645–1654, 2019.

[81] M. I. Razzak, S. Naz, and A. Zaib, "Deep learning for medical image processing: Overview, challenges and the future," in *Classification in BioApps.* Springer, 2018, pp. 323–350.

[82] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[83] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision.* Springer, 2014, pp. 740–755.

[84] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4," *International Journal of Computer Vision*, pp. 1–26, 2020.

[85] A. Fourcade and R. Khonsari, "Deep learning in medical image analysis: A third eye for doctors," *Journal of stomatology, oral and maxillofacial surgery*, vol. 120, no. 4, pp. 279–288, 2019.

[86] J. K. Min, M. S. Kwak, and J. M. Cha, "Overview of deep learning in gastrointestinal endoscopy," *Gut and liver*, vol. 13, no. 4, p. 388, 2019.

[87] C. Matava, E. Pankiv, S. Raisbeck, M. Caldeira, and F. Alam, "A convolutional neural network for real time classification, identification, and labelling

of vocal cord and tracheal using laryngoscopy and bronchoscopy video," *Journal of medical systems*, vol. 44, no. 2, pp. 1–10, 2020.

[88] J. Ren, X. Jing, J. Wang, X. Ren, Y. Xu, Q. Yang, L. Ma, Y. Sun, W. Xu, N. Yang *et al.*, "Automatic recognition of laryngoscopic images using a deep-learning technique," *The Laryngoscope*, vol. 130, no. 11, pp. E686–E693, 2020.

[89] R. Bittner and J. Schwarz, "Inguinal hernia repair: current surgical techniques," *Langenbeck's archives of surgery*, vol. 397, no. 2, pp. 271–282, 2012.

[90] F. Köckerling, D. Jacob, W. Wiegank, M. Hukauf, C. Schug-Pass, A. Kuthe, and R. Bittner, "Endoscopic repair of primary versus recurrent male unilateral inguinal hernias: Are there differences in the outcome?" *Surgical Endoscopy*, vol. 30, no. 3, pp. 1146–1155, 2016.

[91] J. ATGER, "Laparoscopic totally extraperitoneal inguinal hernia repair. twenty-seven serious complications after 4565 consecutive operations," *Rev. Col. Bras. Cir*, vol. 40, no. 1, pp. 032–036, 2013.

[92] A. Meyer, P. Blanc, R. Kassir, and J. Atger, "Laparoscopic hernia: umbilical-pubis length versus technical difficulty," *JSLS: Journal of the Society of Laparoendoscopic Surgeons*, vol. 18, no. 3, 2014.

[93] F. Y. Suguita, F. F. Essu, L. T. Oliveira, L. R. Iuamoto, J. M. Kato, M. B. Torsani, A. S. Franco, A. Meyer, and W. Andraus, "Learning curve takes 65 repetitions of totally extraperitoneal laparoscopy on inguinal hernias for reduction of operating time and complications," *Surgical Endoscopy*, vol. 31, no. 10, pp. 3939–3945, 2017.

[94] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[95] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.

[96] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[97] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[98] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[99] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[100] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[101] R. Kumar and A. Indrayan, "Receiver operating characteristic (roc) curve for medical researchers," *Indian pediatrics*, vol. 48, no. 4, pp. 277–287, 2011.

[102] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, "Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique," *Medical physics*, vol. 43, no. 6Part1, pp. 2821–2827, 2016.

[103] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[104] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, "A deep learning method for classifying mammographic breast density categories," *Medical physics*, vol. 45, no. 1, pp. 314–321, 2018.

[105] B. Zoph and Q. V. Le, “Neural architecture search with reinforcement learning,” *arXiv preprint arXiv:1611.01578*, 2016.

[106] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size,” *arXiv preprint arXiv:1602.07360*, 2016.

[107] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[108] X. Zhang, X. Zhou, M. Lin, and J. Sun, “Shufflenet: An extremely efficient convolutional neural network for mobile devices,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[109] Y. Gong, L. Liu, M. Yang, and L. Bourdev, “Compressing deep convolutional networks using vector quantization,” *arXiv preprint arXiv:1412.6115*, 2014.

[110] L. J. Ba and R. Caruana, “Do deep nets really need to be deep?” *arXiv preprint arXiv:1312.6184*, 2013.

[111] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, “Attention augmented convolutional networks,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3286–3295.

[112] H. Hu, Z. Zhang, Z. Xie, and S. Lin, “Local relation networks for image recognition,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3464–3473.

[113] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, “Stand-alone self-attention in vision models,” *arXiv preprint arXiv:1906.05909*, 2019.

[114] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[115] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[116] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[117] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in neural information processing systems*, vol. 32, 2019.

[118] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 076–10 085.

[119] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[120] J. Guo, Y. Li, W. Lin, Y. Chen, and J. Li, "Network decoupling: From regular to depthwise separable convolutions," *arXiv preprint arXiv:1808.05517*, 2018.

[121] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4055–4064.