A DISSERTATION
SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTER SCIENCE AND ENGINEERING

Study on Computer-Aided Diagnosis in Hysteroscopy Based on Deep Learning

by

ZHAO Aihua

June 2023

The thesis titled

Study on Computer-Aided Diagnosis in Hysteroscopy Based on Deep Learning

by

ZHAO Aihua

is reviewed and approved by:

Chief referee

Senior Associate Professor Date Aug. 11, 2023

Xin Zhu Xin Zhu

Professor Date Aug 11, 2023

Wenxi Chen Wenxi Chen

Professor Date Aug 11 2023

Shigeo Takahashi

Senior Associate Professor Date Aug 28, 2023

Yuichi Yaguchi Yuichi Yaguchi

THE UNIVERSITY OF AIZU

June 2023

Contents

List of Figures

List of Tables

Acknowledgment

Looking back on my journey, I am filled with deep appreciation for those who have provided me with guidance and support along the way.

Foremost, I wish to extend my deepest gratitude to my supervisor, Professor Xin Zhu. His consistent encouragement, guidance, and unwavering support have been fundamental in my research and every life. His support has allowed me to navigate complex challenges and successfully to apply new knowledge. He's shown me what it means to always strive for the best. His influence has shaped my studies and future life.

I extend my heartfelt appreciation to doctors Wenwen Wang and Suzhen Yuan from Tongji Hospital at Huazhong University of Science and Technology, and Professor Wenfeng Shen from Shanghai University. Their generous assistance and support, particularly during our weekly meetings, have been invaluable. Their guidance has illuminated my path, and their insights have helped me navigate through complex medical knowledge, experimental equipment, and methodologies.

My heartfelt appreciation goes out to my friends at the University of Aizu. The experience of learning, working, and engaging in recreational activities with you all over these past few years has been genuinely thrilling. Wishing each and every one of you a future filled with success and prosperity.

The accomplishment of this thesis wouldn't have been possible without the support and assistance of all esteemed professors. Special gratitude is extended to Professors Wenxi Chen, Shigeo Takahashi, and Yuichi Yaguchi. Their expert guidance, contributions to my research topic, and tireless help in navigating through numerous challenges were invaluable.

Abstract

Endometrial polyps (EP), uterine fibroids (UF), endometrial cancer (EC), and atypical endometrial hyperplasia (AEH) are common uterine disorders. A common surgical technique for diagnosing and treating gynecological diseases is hysteroscopy surgery. However, hysteroscopy examinations rely on the subjective judgment of the hysteroscopy surgeon. Insufficient experience of the surgeon may lead to misdiagnosis, resulting in a decrease in the diagnostic accuracy of the lesions and a delay in patient treatment.

Due to their exceptional data processing capabilities, convolutional neural networks (CNNs) have emerged as powerful tools for analyzing medical images. Therefore, applying a computer-aided diagnosis system based on a CNN for uterine lesions in hysteroscopy images can assist physicians in the diagnosis, alleviate their workload, improve diagnostic accuracy, and help shorten surgical treatment time, reducing the risk of surgical complications.

This dissertation introduces a computer-aided diagnosis (CAD) system specifically dedicated to hysteroscopic common uterine lesions. The system proposes the implementation of separate CAD subsystems, each optimized to address the unique features and diagnostic requirements associated with specific lesion types. The entire CAD system consists of three subsystems.

(1) CAD for recognizing UF based on CNN-Transformer hybrid network.

A severe physical health issue for women is UF. Hysteroscopic surgery is an effective way to treat this disease. Considering that the diagnostic requirement for UF does not mandate knowledge of their precise location but rather the identification of their presence in hysteroscopic images, we propose the implementation

of a CAD subsystem based on a classification network to facilitate the diagnostic process. Given the strong ability of a CNN to gain local features and the Transformer architecture to obtain global features, in this CAD subsystem, we hybridize CNN and Transformer networks, introducing learnable parameters, to recognize UF for assisting clinicians. The CAD subsystem was trained by a training dataset composed of 9524 images from 240 patients and 199 healthy subjects. Through an evaluation using a test set composed of 2312 images from 33 patients and 36 healthy subjects, the sensitivity, specificity, accuracy, F_1-score, precision, and AUC of the proposed method are 94.21%, 83.76%, 88.93%, 89.36%, 84.99% and 0.96, respectively. The proposed method outperformed base models including ConvNeXt, Swin Transformer, and Conformer network via testing and comparison, and may be used as a CAD tool for recognizing UF.

(2) CAD for detecting EP based on deep learning.

EP are common gynecological lesions. The standard treatment for this condition is hysteroscopic polypectomy. However, this procedure may be accompanied by misdetection of endometrial polyps. To improve diagnostic accuracy and reduce the risk of misdetection, a CAD subsystem based on YOLOX is proposed to detect endometrial polyps in real-time. Group normalization is employed to improve its performance with large hysteroscopic images. In addition, we propose a video adjacent-frame association algorithm to address the problem of unstable polyp detection. Our proposed CAD subsystem was trained on a dataset of 11,839 images from 323 cases provided by a hospital and was tested on two datasets of 431 cases from two hospitals. The results show that the lesion-based sensitivity of the model reached 100% and 92.0% for the two test sets, compared with 95.83% and 77.33%, respectively, for the original YOLOX model. This demonstrates that the improved CAD subsystem exhibits high sensitivity and may be used effectively as a diagnostic tool during clinical hysteroscopic procedures to reduce the risk of missing EP.

(3) CAD for classifying EC and AEH based on deep learning.

EC is the most common and rapidly increasing female cancer globally. AEH is a precancerous condition of EC. Both EC and AEH are difficult to distinguish from other benign tumors based on their shape. Therefore, we propose a CAD subsystem utilizing the EfficientNet network as a baseline, incorporating the ParNet attention mechanism and class weighting to accurately classify EC/AEH from benign lesions. This study includes 49,556 hysteroscopy images from 1,237 cases as a training set and 3,412 hysteroscopy images from 85 cases as a testing set. AUC, accuracy, sensitivity, specificity, PPV, Kappa, and F_1-score of the proposed method are 0.941, 89.4%, 93.7%, 87.1%, 73.3%, 0.755, and 0.8225, respectively. The CAD subsystem has high sensitivity and may be used as a tool for the diagnosis of EC/AEH.

Chapter 1

Introduction

1.1 Overview

With the development of endoscopic treatment technology, hysteroscopy has become widely used for the diagnosis of uterine diseases. Hysteroscopy is considered to be the gold standard tool for endoscopic visualization of the uterine cavity [3]. Direct visualization of the uterine cavity is made possible by hysteroscopy, a surgical technique frequently used to diagnose and treat gynecological disorders. This enables precise identification of pathological conditions. The use of hysteroscopy can detect and treat common uterine lesions such as uterine fibroids, endometrial polyps, and endometrial cancer. Further, hysteroscopic treatment can increase female fertility. Uterine lesions are detected by hysteroscopy in 10-15 % of women seeking medical treatment for infertility [4].

However, hysteroscopy relies entirely on the subjective judgment of hysteroscopists, and inexperienced and experienced hysteroscopists may be deficient in accurately identifying and diagnosing lesions.

In recent years, the recent rapid development of computer technology has greatly benefited the medical field, particularly in the use of convolutional neural networks (CNN). With the development of computing power, CNNs structure has been deepened and evolved into deep learning networks. Because of their extraordinary data process-

ing capabilities, deep learning networks have developed into potent tools for analyzing medical images. A deep learning network is widely used in computer-aided diagnosis (CAD) systems, and CAD is a method that uses computer algorithms and artificial intelligence techniques to assist in medical diagnosis. CAD systems can analyze medical imaging data such as CT scans, MRIs, and X-rays to provide doctors with information that can assist in disease prediction, disease diagnosis, and lesion detection.

CAD applied in the analysis of hysteroscopic images has shown promising results in the context of gynecological diseases, which present significant challenges to women's health and well-being. Additionally, the use of CAD in hysteroscopy aids in the optimization of surgical techniques. CAD can offer insightful information and support surgical decision-making by examining real-time hysteroscopic images. This not only shortens the total surgical time but also keeps the process from being overly intrusive, which lowers the risks and potential consequences linked to conventional surgical techniques.

The use of CAD in hysteroscopy offers major advantages from an economic perspective in addition to the therapeutic benefits. The use of CAD in hysteroscopic image processing reduces costs for both healthcare professionals and patients by enhancing the precision of diagnosis and streamlining surgical procedures. Reduced hospital stays, fewer surgical complications, and lessening of the use of healthcare resources all contribute to this.

Therefore, we propose a CAD system to automatically detect and classify common uterine lesions under hysteroscopy, including polyps, fibroids, endometrial cancer, and atypical hyperplasia. We propose three specific CAD subsystems to assist in the diagnosis of different lesions with diverse characteristics and diagnostic requirements. We use a large amount of hysteroscopic data from various hospitals to train our CAD to enhance its robustness.

1.2 Hysteroscopic Lesions

1.2.1 Common Uterine Lesions

Due to their prevalence and significant effects on reproductive health, uterine diseases are a significant health concern for women and warrant careful consideration and research. Uterine lesions refer to various pathologies that occur within the uterine region, leading to abnormal uterine positioning. Common uterine diseases include uterine fibroids, endometrial polyps, and endometrial cancer. Figure 1.1 shows various uterine lesions under hysteroscopy.

Uterine Fibroids

Uterine fibroids (UF), also known as uterine leiomyomas, are benign tumors that grow inside the uterus's muscular wall. UF typically resides intramurally within the uterine wall. They commonly exhibit a regular shape, predominantly circular or oval, and may possess a smooth contour or irregular surface indentations. These fibroids can manifest across a wide size spectrum, ranging from minute dimensions of a few millimeters to substantial dimensions measuring tens of centimeters. However, a number of factors, including hormone imbalances, genetic predisposition, and specific growth factors, have been linked to their development. The size of UF can vary, from small seedlings to enormous masses that change the uterus's size and form.

Women's health may be significantly impacted by UF. Over 70% of women encounter UF during their reproductive years, according to the statistics in [5]. Although UF is regarded as a benign tumor, its rapid growth can cause a number of issues and have a negative impact on women's physical health. Heavy or protracted menstrual bleeding, pelvic pain or pressure, frequent urination, and possibly infertility are signs of UF.

For many women, the impact of UF on fertility is a major worry. Fibroids can make it difficult to conceive by interfering with the implantation of a fertilized egg or by affecting the blood flow to the uterus, depending on their size, position, and number.

Furthermore, fibroids might result in pregnancy issues such an increased chance of miscarriage, premature labor, and fetal malpresentation.

In addition, UF can have psychological and emotional impacts on women that result in stress, anxiety, and a lower quality of life. Physical symptoms and potential reproductive effects might cause discomfort and may need proper management and assistance from healthcare professionals.

Therefore, it is essential to correctly diagnose UF and evaluate its effect on women's health, especially those who are attempting to get pregnant. Based on the severity of symptoms, the desire for fertility, and unique patient variables, timely intervention and management techniques, ranging from conservative approaches like medication to surgical options like myomectomy or hysterectomy, might be taken into consideration.

In conclusion, UF is a common illness that significantly affects women. Despite being benign, their fast development may be harmful to physical well-being and fertility. Healthcare professionals must be aware of the effects of UF and offer suitable therapeutic strategies to reduce symptoms, maintain fertility when desired, and enhance the general well-being of women with this illness [6].

Endometrial Polyps

Endometrial polyps (EP), benign tumors originating from the excessive growth of endometrial glands, stroma, and blood vessels within the uterine lining, exhibit a diverse range of characteristics. They can affect women across all age groups, with a peak incidence noted among those aged 40 to 49 [7, 8].

The shapes of EP can exhibit considerable variation. Common shapes of endometrial polyps include:

(1) Elongated: taking on an elongated form, resembling an extended ellipse.

(2) Spherical: characterized by a round or nearly round shape.

(3) Oval: displaying an elliptical shape, with varying ratios of the long axis to the short axis.

(4) Tapered: featuring a wider base that gradually narrows toward the tip.

(5) Irregular: endometrial polyps may manifest irregular shapes, with uneven or lobulated edges.

Moreover, the size and shape of EP can differ among individuals, further contributing to their morphological diversity. The shape and structure of polyps play a crucial role in their impact on women's health, particularly larger polyps that can distort and alter the normal uterine structure by occupying a significant portion of the uterine cavity.

EP can elicit various symptoms and complications, most notably abnormal uterine bleeding. This can manifest as heavy or prolonged menstruation, irregular bleeding between cycles, or postmenopausal bleeding. Pelvic pain, ranging from mild discomfort to severe cramping, can also occur due to the presence of polyps. Moreover, these polyps can impede successful embryo implantation and contribute to recurrent miscarriages, thereby affecting fertility outcomes.

While EP are generally considered benign tumors, it is important to acknowledge the potential for malignant transformation. Reported rates of malignant transformation within endometrial polyps range from 0% to 13% [9,10]. Thus, a thorough evaluation is necessary to assess the possibility of malignancy.

Endometrial Cancer and Atypical Endometrial Hyperplasia

Endometrial cancer (EC) is a common gynecologic malignancy primarily affecting the endometrium [11]. Its incidence has been increasing rapidly, especially in high-income countries, where it ranks as the fourth most common cancer in women [12,13]. In 2020, 417,367 new cases and 97,370 deaths worldwide were attributed to EC [14]. Although EC generally has a good prognosis, high-grade cancers are at higher risk of recurrence and may present at advanced stage or with metastatic disease, which presents a major challenge to treatment [15].

Atypical endometrial hyperplasia (AEH) is considered a precancerous lesion of the endometrium. After hysterectomy, EC was diagnosed in 27% to 52% of patients with

preoperative AEH [16]. The risk of endometrial cancer in patients with AEH during hysterectomy is approximately 40% [17]. Given the rapid progression of these lesions, accurate diagnosis of EC and AEH is critical for early intervention.

EC is the abnormal proliferation of endometrial cells, and malignant transformation to form a glandular structure, usually caused by genetic factors and excess estrogen. Depending on the growth pattern and extent of the tumor's spread. EC can vary in shape from massive, polypoid, and invasive, to diffuse. EC can cause a range of symptoms, including abnormal uterine bleeding, pelvic pain, and abdominal discomfort. If left untreated or diagnosed at an advanced stage, it causes metastasis and significantly reduces overall survival. AEH typically presents with glandular congestion, complex branching patterns, and increased cellularity. As a premalignant lesion, AEH has the potential to develop into EC if not managed effectively. Therefore, early diagnosis and appropriate treatment are crucial to improve patient prognosis and outcomes.

1.2.2 Hysteroscopy Surgery

Hysteroscopy is currently considered the gold standard for endoscopic visualization of the uterine cavity. Hysteroscopy can introduce micro-surgical instruments such as scissors, coagulation forceps, and sampling forceps for surgical treatment. As shown in Figure 1.2. The hysteroscopy can enter the uterine cavity through the vagina, and transmit the image of the internal conditions of the uterine cavity to the display screen through the optical fiber, thereby providing a clear view for the doctor and helping the doctor to observe and check the endometrium, uterine cavity and nearby tissues situation. Hysteroscopic surgery can be used to diagnose and treat a variety of uterine-related diseases, including endometrial polyps, uterine fibroids, and endometrial cancer. With the development of rapid endoscopic technology, hysteroscopy has become an effective minimally invasive procedure, which causes less trauma than traditional open surgery, less postoperative pain, faster recovery time, and less risk of complications.

However, hysteroscopy relies solely on the subjective judgment of hysteroscopists, which may introduce inherent variability and potential errors into the diagnostic pro-

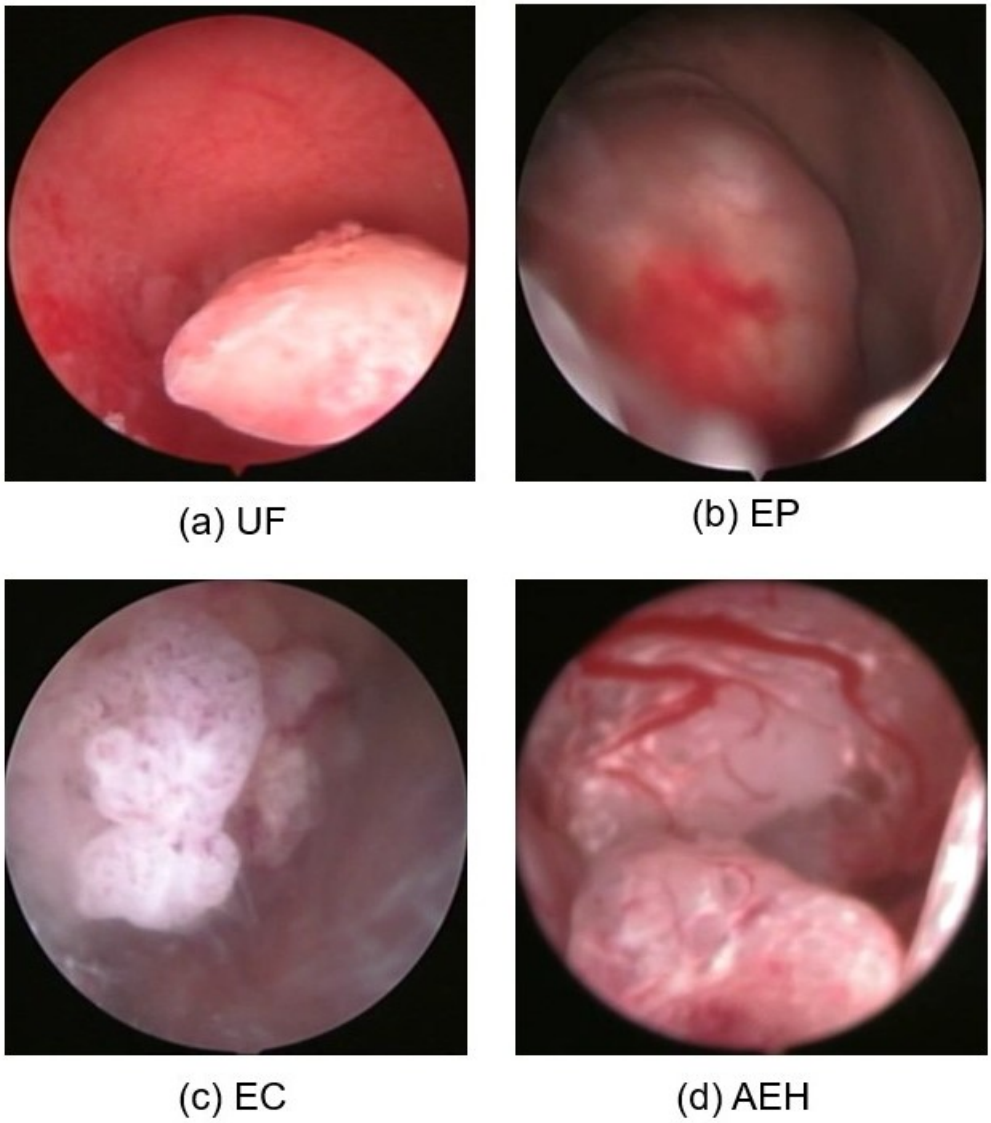


Figure 1.1: Hysteroscopic images of uterine lesions. (a) Uterine fibroids, (b) Endometrial polyps, (c) Endometrial cancer, (d) Atypical endometrial hyperplasia

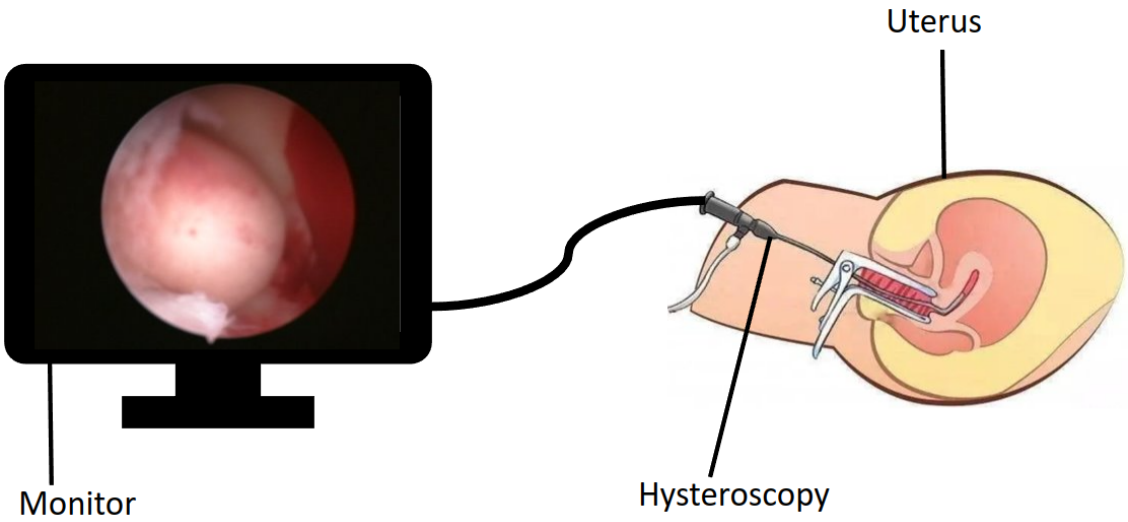


Figure 1.2: Hysteroscopic surgery procedure.

cess [18]. Complications of hysteroscopic surgery—such as intraoperative bleeding, uterine perforation, peripheral organ injury, water intoxication, and intrauterine adhesions—should be of great concern [19]. Water intoxication can cause systemic toxicity and even death, resulting from the fluid load during surgery. Experienced hysteroscopists play an important role in reducing these intraoperative and postoperative complications and can make an initial prediction of the extent of the disease. Inexperienced hysteroscopists may face challenges in accurately identifying and interpreting abnormalities, resulting in decreased diagnostic accuracy. This, in turn, can lead to delays in initiating timely and appropriate treatment for patients, leading to suboptimal outcomes and potential economic impact.

Therefore, interventions focused on enhancing the objectivity and precision of hysteroscopy, especially by incorporating computer-aided diagnosis systems, are urgently needed. These interventions are designed to mitigate the challenges posed by limited clinical expertise, thereby reducing the risks associated with misdiagnosis, treatment delays, and financial impact.

1.3 Development of Deep Learning Methods

1.3.1 Convolutional Neural Networks

An artificial neural network, or simply a neural network, is a computing model that mimics the structure and functions of biological neural networks for distributed parallel information processing. Artificial neural networks consist of an input layer, one or more hidden layers, and an output layer. The input layer receives the input data, the hidden layer processes intermediate information, and the output layer delivers the final result. The neural network comprises a large number of neurons and connections between neurons and is capable of adaptively performing nonlinear transformations of the input data using activation functions. The connections between neurons are accompanied by weights, and the neural network uses gradient descent to adjust these weights to minimize the value of the loss function, i.e., to reduce the difference between the generated and real values of the neural network. Each neuron has an activation function that is used to perform a nonlinear transformation of the input signal.

A CNN is a special kind of artificial neural network whose most important feature is the introduction of convolution and pooling operations. In each region of the image, the process of replacing the pixel value of one point with the weighted average of the pixel values of its surrounding points as a new value by a linear transformation is known as the convolution process. The process of multi-layer convolution is the process of transforming pixel values layer by layer, and the neural network learns the weights needed to transform each local pixel region value during the training process. Figure 1.3 shows the developments of CNN.

Around 1980, Japanese scientist Kunihiko Fukushima proposed the neurocognition model [20], which was a neural network with a deep structure, consisting of a network layer of S cells for extracting local features and a network layer of C cells for abstraction and fault tolerance, and these two network layers are alternately combined to realize the convolutional and pooling layers in the current CNN operations. Kunihiko Fukushima used the model for recognizing handwritten characters, and the model was seen as a

precursor to modern CNNs. Subsequently, Yann ec at. [21] introduced the backpropagation algorithm into the Neocognitron framework, significantly reducing the number of parameters through the principle of weight sharing and feature mapping. This led to the proposal of LeNet-5 [22], a convolutional neural network that utilizes gradient learning, in 1998. This innovative network was capable of recognizing handwritten digits with an error rate of less than 1% and was subsequently implemented in virtually all postal systems across the U.S. for digit recognition. The development of LeNet-5 played a pivotal role in laying the foundation for future advancements in CNN.

The advancement of computational devices led to the proposition of the AlexNet model in 2012 [23]. This model harnessed the parallel computing prowess of GPUs, implementing a deeper network structure and deploying the Rectified Linear Unit (ReLU) activation function. The latter effectively addressed the vanishing gradient problem during training, accelerating convergence. In addition, the Dropout technique was utilized to curb overfitting in the model. The model was trained on the expansive ImageNet image dataset [24]. The triumph of AlexNet ignited a revolution in the realm of neural networks, signifying the ascent of deep learning in the field of computer vision.

In 2014, VGGNet was proposed [25]. The model had a simple and deep network structure, introduced the Inception module, and extracted multi-scale features by using convolutional and pooling kernels of different sizes in parallel. Kaiming He et al. proposed ResNet in 2015 [26], which introduced residual connectivity to alleviate the gradient disappearance and gradient explosion problems caused by increasing depth in deep network training. This design allows the network model to be deeper, further improving the performance of image classification. In 2017, the lightweight neural network model MobileNet [27] was proposed to achieve high efficiency in real-time image recognition and processing on mobile devices. MobileNet used depthwise separable convolution to decompose a standard convolution into two smaller operations: depthwise convolution and pointwise convolution, reducing the computational effort and number of parameters. EfficientNet [2] was proposed in 2019, and the model proposes a scaling strategy that combines depth, width, and resolution to achieve a more

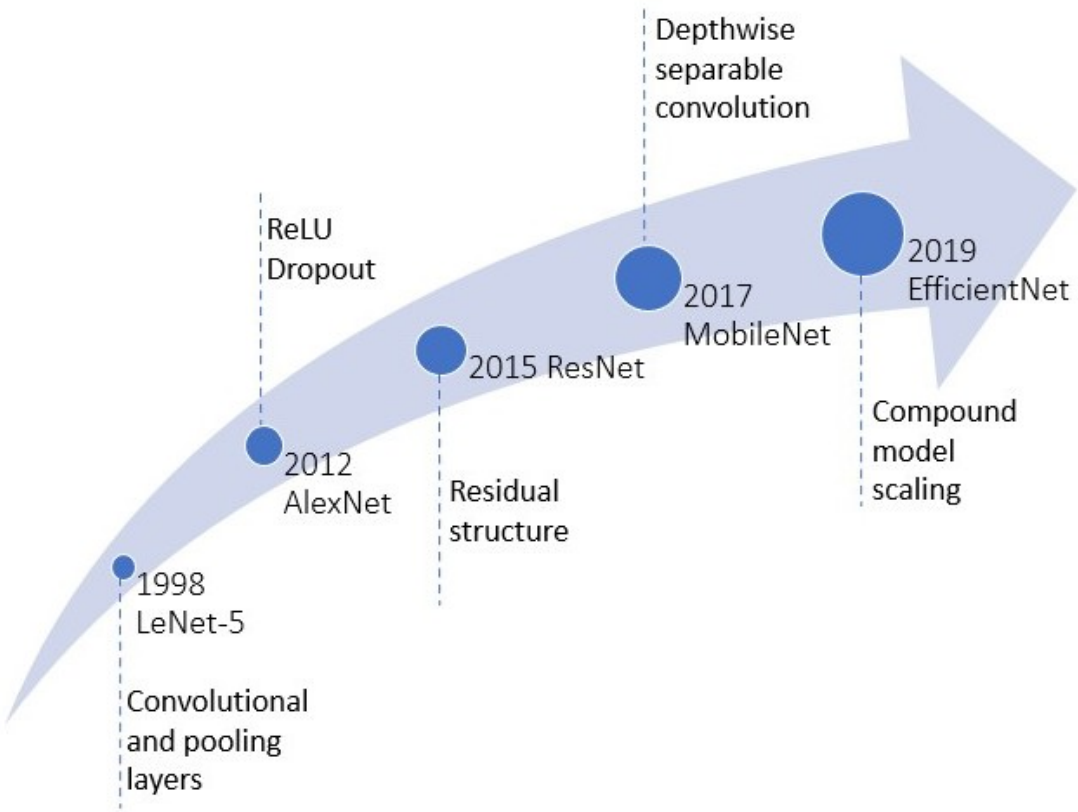


Figure 1.3: The developments of CNN.

efficient network structure.

1.3.2 Transformer

The Transformer model [28] is a deep neural network model using the self-attention mechanism, originally proposed by Google in 2017. To achieve better parallelization and shorter training time, the model uses the attention mechanism instead of the recurrent structure in recurrent neural networks. The model was initially used to solve sequence modeling problems in natural language processing, and now Transformer is also widely used for image tasks [29]. Figure 1.4 shows the developments of Transformer.

Ramachandran et al. [30] proposed a fully attentional vision model, which uses Attention-based and relative position embedding modules to completely replace the convolutional modules. The model used a local attention approach, i.e., each pixel is computed with only a few pixels around it for attention, which effectively reduced the

computational overhead. The model achieved excellent results on both image classification and detection tasks compared to traditional CNN models. The model achieved a 12% FLOPS reduction and 29% parameters reduction over the ResNet model for the ImageNet classification task. On COCO object detection, it achieves a 39% FLOPS reduction and 34% reduction over the baseline model RetinaNet parameters.

Carion et al. [31] proposed the DERT model, which proposed the idea of considering object detection as a direct set prediction problem. The architecture of the DERT model consists of CNN, Transformer encoder, and Transformer decoder. Initially, the input image is processed through the CNN to generate a feature map, which is subsequently fed into the encoder for encoding. Following encoding, the map is sent to the decoder to produce a predefined quantity of prediction frames. The frames then undergo processing via a feed-forward network to generate the final result prediction. Prediction frames that align with the ground truth are retained, while all others are dismissed.

Chen et al. [32] introduced the iGPT model in 2020, an innovation that applies unsupervised pre-training to large-scale image data using the GPT-2 model structure. The model overlooks the two-dimensional structural information of an image and reshapes it into a one-dimensional sequence for input into the Transformer, subsequently executing the next pixel prediction tasks. In light of the substantial input size typically involved in image classification, the authors employed image downsampling and k-means clustering to further mitigate the input volume. Furthermore, the authors contrasted two pre-training optimization strategies - auto-regressive and BERT objectives.

The ViT model was proposed in 2021 by Dosovitskiy et al. [29], adhering closely to the original structure of the Transformer model for image classification tasks. The ViT model segments the image into multiple patches, subsequently linearly mapped to attain a fixed dimensional representation for input into the Transformer encoder. The model utilizes a supervised approach for ongoing training. A special classification token is appended at the beginning of the sequence for image classification. ViT achieves impressive results by training on extensive datasets (14M-300M images) without incorporating the translation invariance and two-dimensional spatial relationships typical

of traditional convolutional neural networks. Pre-trained on publicly available datasets such as ImageNet-21K or JFT-300M [33], ViT equals or surpasses SOTA benchmarks for multi-sample classification, though its accuracy falls short of ResNet when trained on medium-sized datasets like ImageNet.

Han et al. [34] argue that the ViT model's patch segmentation lacks granularity to extract features from objects of varying scales and locations within the image. As a remedy, they presented the TNT model which further divides the patch into smaller 4×4 segments, referred to as visual words. The model enhances feature representation by examining the relationships between patches through a self-attentive mechanism.

The Pyramid Vision Transformer (PVT) model, proposed by Wang et al. [35], was designed to address the ViT model's limitations for dense detection tasks. The reduced output resolution and increased computation and memory overhead associated with the growth of sequence length after ViT image patch division prompted the incorporation of pyramidal convolutional neural network concepts. The PVT model balances the capture of more fine-grained information with reduced computational overhead by modifying the resolution in a layer-by-layer manner. The model is segmented into four stages, each encompassing image block embedding and Transformer. The output resolution decreases progressively with each stage, facilitating the capture of information at diverse scales and compatibility with various pixel-level image tasks. The PVT model demonstrates notable performance on multiple datasets.

Liu et al. [36] proposed the Swin Transformer model, which similarly to the PVT model, segments the image into smaller blocks and reduces resolution in a layer-by-layer approach to capture multi-scale features. However, in contrast to traditional convolutional neural networks, Swin Transformer employs a local attention mechanism, dividing image blocks into windows, with attention computation only occurring within each window rather than across the whole image. To promote interlayer information interaction, the model introduces the concept of "Shifted Windows." The window's position is fine-tuned horizontally and vertically by 2 patches in each layer, according to the offset of the previous layer. This strategy facilitates interaction between image

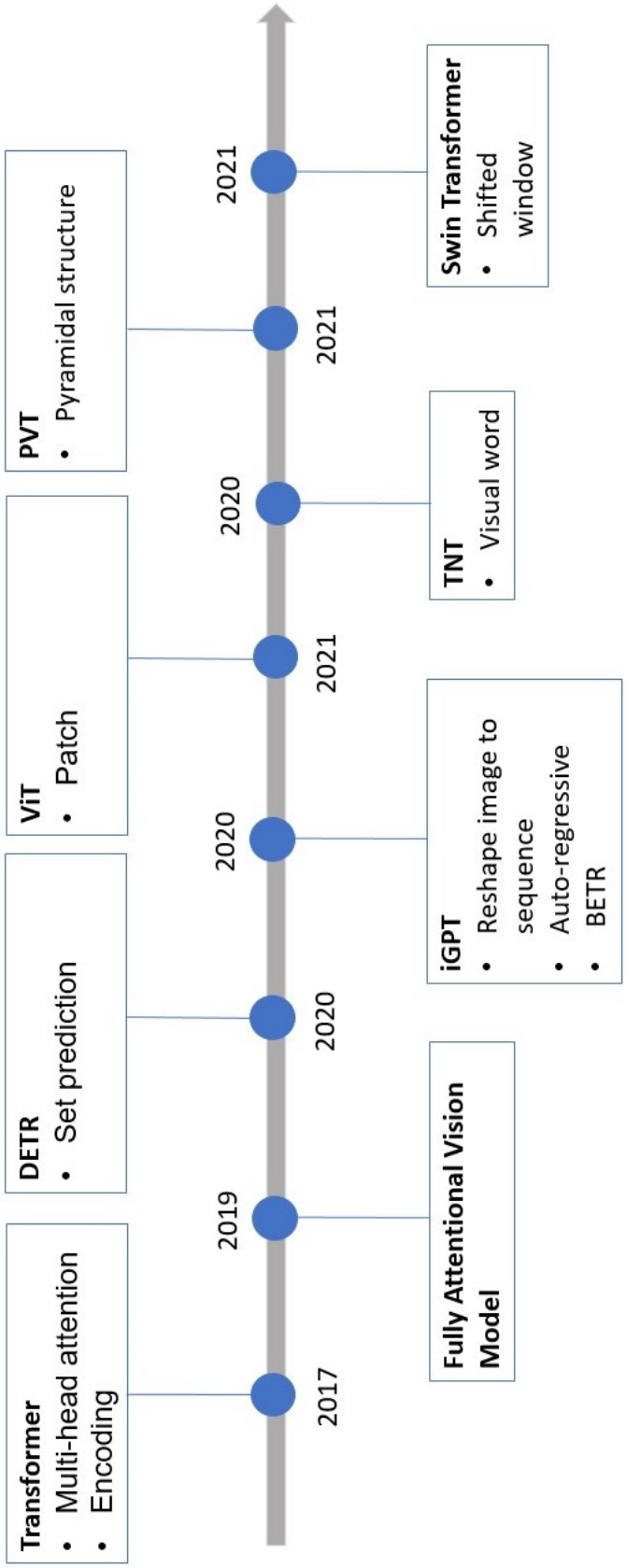


Figure 1.4: The developments of Transformer.

blocks of different windows, enhancing the model's perceptual field and promoting feature propagation across windows.

CNN is a hierarchically structured neural network containing multiple convolutional and pooling layers, as well as fully connected layers for classification, through which local features of the image are extracted. The convolution layer performs a convolution operation on the image by means of a sliding window to extract local features and reduces the spatial dimension of the feature map by means of a pooling layer. However, when operations such as pooling are performed, the relationships between objects in the image may be lost, the CNN ignores the overall and local associations of features in the image, and the CNN lacks a comprehensive understanding of the image.

In contrast, Transformer is a structure consisting of multiple self-attentive layers and feedforward neural network layers without convolution and pooling layers. It performs global feature attention on the whole input sequence through the self-attentive mechanism and feature transformation through the feedforward neural network to achieve effective modality fusion capability. It is able to capture the relationships at different positions in the sequence and achieve multi-angle feature representation and combination, through a multi-headed self-attention mechanism. This multi-head mechanism can capture the rich information in the input data and interact and fuse them at different levels, thus improving the representativeness and performance of the model. This global feature attention enables Transformer to better capture long-distance dependencies and global contextual information in images, which can be noticed in each subspace and can capture richer feature information.

1.4 Computer-Aided Diagnosis System based on CNN

The rapid advancement of CNN has drawn increasing scholarly attention toward their potential integration in the medical field, specifically in the development of Computer-Aided Diagnosis (CAD) systems. CAD has emerged as a potent tool in healthcare, enabling medical professionals to enhance patient care via precise and efficient diag-

nostic processes. This section delves into the incorporation of CNN into CAD systems, elucidating their potential in augmenting diagnostic accuracy and therapeutic outcomes.

CNN, an integral component of deep learning, has shown tremendous promise in image recognition tasks, due to its ability to automatically and adaptively learn spatial hierarchies of features directly from data. This aptitude is particularly beneficial in the context of medical imaging, where nuanced patterns often elude traditional image-processing techniques. This capability has unlocked new possibilities for CAD systems in various medical imaging domains, including mammography, dermatology, radiology, pathology, and endoscopy. Thus, the integration of CNN in CAD systems has opened up avenues for more accurate and precise medical diagnoses, promoting improved patient care.

Emerging research accentuates the effectiveness of CNN-based CAD systems across diverse medical domains. For instance, in mammography, CNN-based CAD systems have shown significant promise in detecting early-stage breast cancers, enhancing diagnostic precision, and promoting early intervention [37–42].

In the field of dermatology, CNNs have been successfully employed in CAD systems to classify skin lesions, even achieving dermatologist-level accuracy in some instances [43–52]

In the domain of radiology, deep learning models have been incorporated in CAD systems for the analysis of chest X-ray images, aiding in the detection of diseases like pneumonia and tuberculosis [53–59].

Similarly, in pathology, CNN-based CAD systems have been developed for the analysis of histopathological images, enabling the classification of different types of cancerous and non-cancerous cells [60–63].

Further, in endoscopy, CAD systems are widely used to diagnose colonoscopy images, perform detection and classification of colorectal polyps, etc [64–74].

Particularly, CAD has been integrated into the field of uterine disease diagnosis by several researchers, demonstrating its potential as a valuable tool in this domain. Kundu et al. propose CAD based on Inception v3, DenseNet-161, and ResNet-34 for classi-

fying single-cell and slide images of cervical cancer [75]. Zhao et al. present a Multi-Modality Ovarian Tumor Ultrasound (MMOTU) dataset and a feature alignment-based architecture called DS2Net for unsupervised cross-domain semantic segmentation of 2D ultrasound images, aiming to address the lack of research on exploring the representation capability of multi-modality ultrasound ovarian tumor image [76]. Wang et al. utilize a CAD combining a deep edge-aware network and marker-controlled watershed algorithm to extract bubble parameters from hysteroscopy images, providing a reference for automatic bubble removal devices in hysteroscopic surgery. [77]. Song et al. propose a CAD aimed to test the feasibility of deep learning-based classification using whole slide images (WSIs) for subtyping cervical and endometrial cancers and determining the origin of adenocarcinomas [78].

In view of the compelling scholarly evidence available, our research explores the feasibility of employing CAD systems in the diagnosis of uterine lesions from hysteroscopy images. We envision that the integration of CNNs in such CAD systems could potentially revolutionize gynecological diagnostics, enhancing the accuracy and efficiency of detecting and differentiating between various types of uterine lesions. The promising results from preceding research add credence to our venture and provide a robust scientific foundation for our exploration. As we continue to refine our model and conduct more rigorous evaluations, we hope that our research will contribute to the ongoing advancements in the application of CNNs within CAD systems, ultimately benefiting patient outcomes and the wider medical community.

1.5 Computer-Aided Diagnosis System in Hysteroscopy

Deep learning has been widely applied in medical fields, such as disease diagnosis and prediction. For the diagnosis of uterine lesions, most deep learning-based CAD systems focus on pathological images, ultrasound images, and MRI images. Hysteroscopy is an important tool for diagnosing uterine diseases and is widely used clinically. However, at present, CAD diagnostic systems for hysteroscopic images are relatively scarce.

The diagnosis of hysteroscopy is subjective and relies on experienced hysteroscopists. Due to the diverse shapes of uterine lesions and large individual variability, inexperienced hysteroscopists can easily miss them.

Given the above, we believe it is necessary to develop a CAD system specifically for the diagnosis of hysteroscopic images to assist physicians in clinical diagnosis. The CAD system can train on a large number of hysteroscopic images through deep learning, thereby accurately identifying abnormalities in the images. This is more accurate than simply relying on a doctor's experience, especially in identifying small lesions and those that are difficult to recognize. Figure 1.5 illustrates the dissertation's motivation.

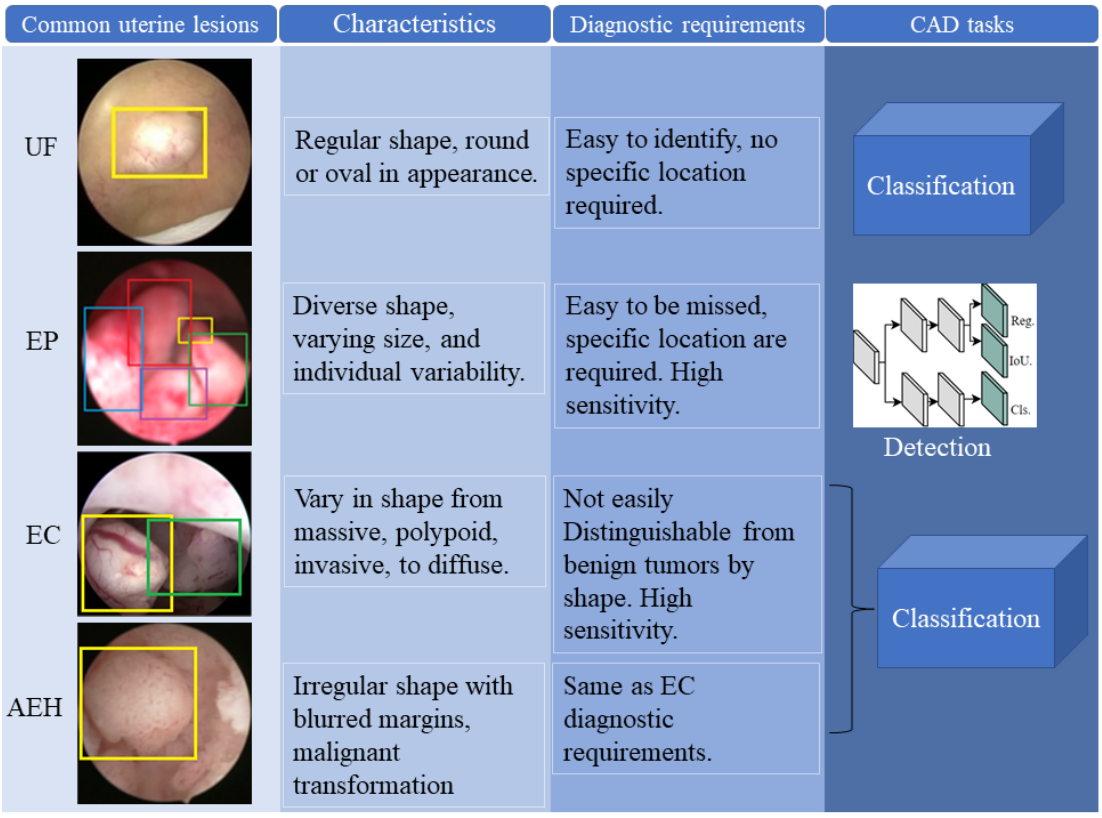


Figure 1.5: The dissertation's motivation. UF: uterine fibroid, EP: endometrial polyp, EC: endometrial cancer, AEH: atypical endometrial hyperplasia. The areas boxed out are the lesions in hysteroscopic images.

During the development of the CAD system, we need to consider the following factors:

(1) Different lesions show large shape differences and individual variability. For instance, UF is regular in shape, but EP vary in shape and size, and the shapes of

EC and AEH are complex.

(2) Different hospitals use different types of hysteroscopic equipment, leading to variations in the lighting and angles of hysteroscopic image acquisition.

(3) Different hysteroscopists have different image acquisition habits.

(4) Different diagnostic objectives are proposed for different lesions when applying the CAD system in clinical hysteroscopic surgery. For instance, UF has a regular shape, making them easy to identify. In clinical diagnosis, the existence of UF can be determined without specific location information. EP has irregular shapes, sizes, and individual variability, making them hard to visualize during hysteroscopic surgery. Junior hysteroscopists may unnecessarily prolong the surgery time or miss polyps, so a highly sensitive CAD system is required. EC and AEH are difficult to distinguish from other benign tumors based on their shape. Insufficient diagnosis of EC leads to inappropriate surgical risks and delays the best time for patient treatment. Therefore, the CAD system must have a high ability to distinguish benign and malignant lesions.

The above factors make the hysteroscopic images complex and complicate the development goals of our CAD system, presenting difficulties for the training of deep learning models. Therefore, we need to evaluate specific CAD subsystems for different lesions according to different diagnostic objectives clinically and improve the deep learning model according to the application of the system to adapt to diagnostic objectives.

The ultimate goal of this dissertation is to propose a deep learning-based CAD system to assist in diagnosing uterine lesions under hysteroscopy. To achieve this goal, the following research is proposed:

(1) In clinical diagnosis, since UF is regular and easy to recognize, the existence of UF can be determined without specific location information. A deep learning-based classification model can meet the needs of auxiliary diagnosis of UF, it

consumes less computation and memory, and the training and inference time is shorter. Although CNN obtains local features, it lacks a comprehensive understanding of the image. The Transformer can effectively collect global information but requires a long training time to pay attention to local features. Therefore, we propose a CAD subsystem based on a hybrid network to identify UF.

(2) EP has irregular shapes, sizes, and individual variability, making them easy to miss during hysteroscopic surgery. Therefore, an object detection network is needed to assist physicians in diagnosis. We propose a highly sensitive CAD subsystem based on YOLOX to detect EP.

(3) EC and AEH have complex shapes and are not easily distinguishable from other benign tumors by their shape, and AEH tends to become malignant. If the diagnosis is not timely, it will delay the best treatment opportunity for the patient. Therefore, we improved a novel and efficient EfficientNet as the base model of the CAD subsystem for classifying EC and AEH.

In summary, we have taken into account the distinctive features of different lesions and put forth three specific CAD subsystems trained with a large amount of hysteroscopic data from different hospitals. We will develop CAD subsystems for diagnosing endometrial hyperplasia and cervical ectropion in the future. Our future endeavors involve amalgamating these subsystems into a holistic CAD hysteroscopy system platform, thereby facilitating its utilization by healthcare professionals and enabling its integration into telemedicine services or remote healthcare systems. The use of a CAD system can quickly analyze hysteroscopic images, improving the efficiency of diagnosing uterine lesions. The CAD system helps physicians discover potentially missed lesions, reducing misdiagnosis and missed diagnosis. In addition, for inexperienced junior physicians, the CAD system can serve as a learning tool to help them familiarize themselves with and understand various hysteroscopic lesions.

1.6 Main Contributions

This dissertation presents a CAD system for diagnosing common uterine lesions under hysteroscopy, proposing separate CAD subsystems for specific lesions. Figure 1.6 illustrates the schematic system architecture of the CAD system proposed in this dissertation. Our future work involves developing CAD subsystems for diagnosing endometrial hyperplasia and cervical ectropion. This dissertation's contributions can be divided into three main areas:

(1) CAD for recognizing UF based on CNN-Transformer hybrid network.

In order to meet the diagnostic requirement of detecting UF in hysteroscopic images, it is not essential to precisely locate the fibroids but rather to accurately identify their presence. To address this, we propose the implementation of CAD subsystem based on a classification network. This subsystem combines the complementary capabilities of CNN and Transformer architectures, leveraging CNN's proficiency in capturing local features and Transformer's ability to capture global features. By incorporating learnable parameters, the hybridized CAD subsystem effectively recognizes UF, assisting clinicians in their diagnostic process.

(2) CAD for detection of EP based on deep learning.

To enhance the diagnosis accuracy of EP and minimize the potential for misdetection, a CAD subsystem utilizing the YOLOX model is introduced. To further optimize performance, the CAD subsystem incorporates the group normalization method. Moreover, to address the challenge of unstable polyp detection, a novel video adjacent frame association algorithm is proposed. The improved CAD subsystem demonstrates remarkable sensitivity and holds significant potential as a reliable diagnostic tool during clinical hysteroscopic procedures, effectively reducing the risk of overlooking EP.

(3) CAD for EC and AEH based on deep learning.

We propose a CAD subsystem for the accurate classification of EC/AEH from

benign lesions. The subsystem utilizes the EfficientNet network as a baseline, augmented with the ParNet attention mechanism and class weighting method. The integration of these components enhances the sensitivity of the CAD subsystem, positioning it as a valuable tool for the precise diagnosis of EC/AEH.

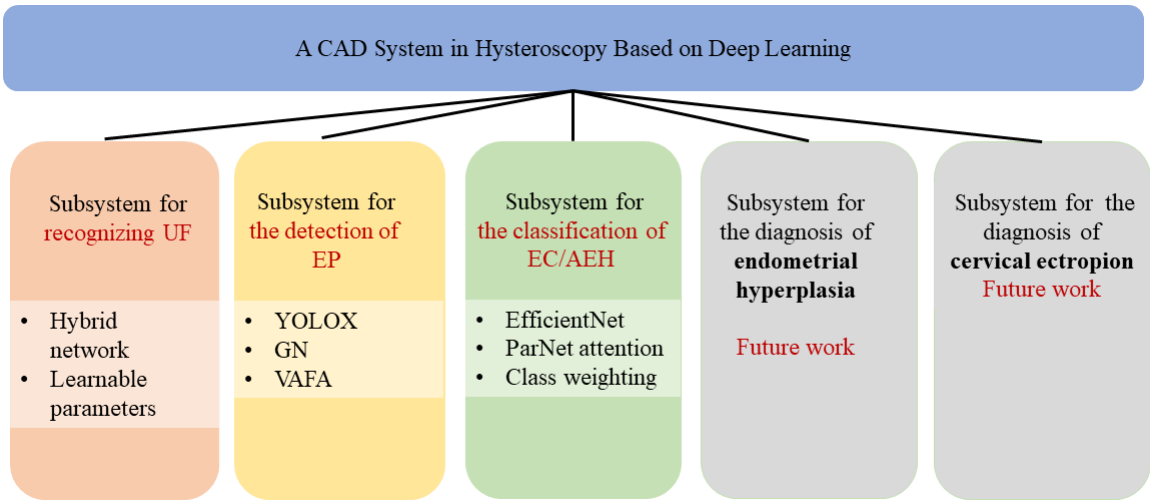


Figure 1.6: The dissertation's motivation. UF: uterine fibroid, EP: endometrial polyp, EC: endometrial cancer, AEH: atypical endometrial hyperplasia.

1.7 Dissertation Outline

The dissertation is primarily divided into five chapters. The dissertation outline is depicted in Figure 1.7.

Chapter 1 serves as an introductory section within this dissertation, aiming to establish the research background and elucidate its significance. We present a thorough description of the CAD system in hysteroscopy diagnosis and research contributions. Additionally, this chapter introduces common uterine lesions, offering a comprehensive understanding of their characteristics and manifestations. Moreover, the chapter delves into the historical evolution of CNN and Transformer, shedding light on their respective advancements and milestones. This elucidation aids in contextualizing the subsequent research methodology and approaches employed within the study. Furthermore, the chapter extensively explores the development and applications of computer-aided diagnosis systems, illuminating their growth, advancements, and practical implementations.

Chapter 2 introduces a CAD subsystem based on a novel deep learning approach for the recognition of uterine fibroids in hysteroscopy images. By leveraging a hybrid network structure that combines CNN and Transformer models, CAD subsystem effectively captures local and global features. The Feature Coupling Unit is enhanced to dynamically focus on more favorable features.

Chapter 3 introduces a CAD subsystem based on YOLOX for the detection of endometrial polyps. The proposed method improves detection performance through the incorporation of group normalization in the YOLOX model. Additionally, a video adjacent-frame association algorithm is implemented to address unstable polyp detection. This CAD subsystem represents the pioneering application of deep learning for endometrial polyp detection in hysteroscopy images, offering significant advancements in the field.

Chapter 4 underscores the imperative for the CAD subsystem in hysteroscopy. The proposed method is based on the EfficientNet-B0 architecture, which introduces a standardized approach for scaling and balancing network dimensions, thereby achieving enhanced performance while maintaining computational efficiency. Integration of the ParNet Attention module within the EfficientNet-B0 architecture yields significant advantages over the traditional SE attention module. The ParNet Attention module incorporates the Skip-Squeeze-Excitation (SSE) block. By replacing the Squeeze-and-Excitation (SE) module with the ParNet Attention module, the model achieves a more nuanced feature representation, improving its comprehension of hysteroscopy images without imposing additional complexity. Furthermore, we use the class weighting method to address the issue of the imbalanced distribution of images across different categories.

Chapter 5 concludes the dissertation by emphasizing our contributions, discussing the limitations of our CAD system, and offering future research plans.

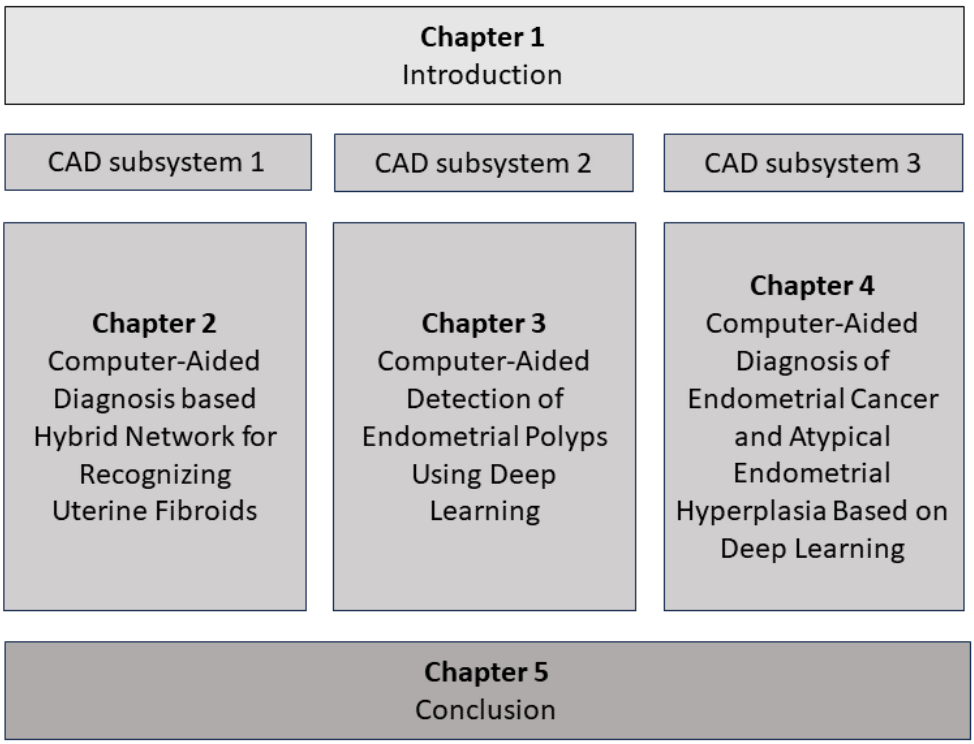


Figure 1.7: The outline of the dissertation.

1.8 Publications

The following papers have been published in peer-reviewed journals and conferences. The majority of the results from Chapters 2, 3, and 4 have been published in these works.

Journals

(1) Zhao, X. Du, S. Yuan, W. Shen, X. Zhu, and W. Wang, "Automated detection of endometrial polyps from hysteroscopic videos using deep learning," Diagnostics, vol. 13, no. 8, 2023.

Conferences

(1) Zhao, X. Du, S. Wang, W. Wang, S. Yuan, W. Ma, W. Yan, W. Shen, and X. Zhu, "A cnn-transformer hybrid network for recognizing uterine fibroids," in 2022 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), 2022, pp. 1-4.

(2) Zhao, X. Du, W. Wang, W. Ma, S. Wang, and X. Zhu, "Deep Learning for Diagnosis of Endometrial Cancer and Atypical Endometrial Hyperplasia," in 2023 IEEE the 4th International Conference on Pattern Recognition and Machine Learning (PRML 2023), 2023, pp.1-6.

Chapter 2

Computer-Aided Diagnosis based Hybrid Network for Recognizing Uterine Fibroids

2.1 Introduction

According to the statistics in [5] over 70% of women suffer from UF. UF is a benign tumor, but its rapid growth can harm women's physical health and potentially impair their ability to conceive [6]. The current endoscopic visualization golden standard for UF diagnosis is hysteroscopy. UF typically exhibits a regular shape, making it easily identifiable by an experienced hysteroscopist during a clinical examination.

However, the clinical experience of hysteroscopists is a major factor in the accurate diagnosis of UF, making it a rather subjective process. Unlike other conditions, the presence of UF can be ascertained without relying on specific location information. This characteristic provides an opportunity to leverage CAD systems for the classification of UF. By harnessing the capabilities of such systems, accurate and efficient identification of UF can be achieved, assisting healthcare professionals in the diagnostic process. The utilization of a computer-aided system for UF classification diminishes the subjective nature of diagnosis and enhances objectivity, contributing to improved clinical decision-

making and patient care.

The deep learning-based method has demonstrated strong data analysis capabilities, which may assist doctors in diagnosis and reduce the subjectivity of diagnosis results. Current deep learning methods can be divided into two categories: CNN-based and Transformer-based methods.

CNN was proposed by Yann LeCun in 1998, which solved the problems of high processing cost, low efficiency, and inaccuracy in traditional algorithms. The CNN network simplifies the traditional image processing process by imitating the visual perception mechanism of the organism and enables the network to efficiently extract the features of the image and reduce the feature dimension through the convolutional and pooling layers. In general, convolutional neural networks have the ability to represent learning and can classify input information according to the spatial hierarchy of learned patterns. Therefore, CNNs are widely used in various computer vision tasks such as image classification, object detection, and image segmentation. Among these tasks, the most critical part is the feature extraction of the target object. CNN is significantly better at extracting features than traditional algorithms; however, because of its translation invariance, the relationship between objects in the image may be lost during operations like pooling, which neglects the overall and local associations of features in an image. As a result, CNN lacks a comprehensive understanding of the image and is unable to comprehend all features. Therefore, it is impossible to establish structural connections globally.

The transformer achieves effective modal fusion capabilities by using a multi-head self-attention technique. It can efficiently gather global information, notice the information of various subspaces, and capture richer feature information. To provide the model with structural information of the image, the visual transformer first divides an image into fixed-size patches and then adds positional embeddings to each patch. Compared with CNN, the visual transformer has a wider receptive field, making it easier to obtain global features such as the relationship between image components and the spatial relationship between objects. However, the transformer in the field of computer vision

also has significant disadvantages. First, it requires an extremely long training time to focus on local areas of an image, and the extraction of local features is not as good as CNN. Secondly, due to its high computational complexity, there is still a gap between the computational complexity of the transformer and the excellent CNN.

Therefore, we proposed a CAD subsystem that combined the advantages of CNN in effectively obtaining local features and the advantages of the visual transformer in capturing global features for the recognition of UF using hysteroscopy images. In this study, we adopted the hybrid network structure Conformer [1] and improve its Feature Coupling Unit to be suitable for the recognition of UF based using hysteroscopic images. We called the CAD subsystem a UF classification system (UFCs), to learn the local features and global representations of UF images to better assist doctors in diagnosis and reduced the misdiagnosis rate.

The contributions of this chapter are as follows.

- We improved the Feature Coupling Unit to weight the features of CNN and transformer so that the UFCs had the ability to dynamically focus on more favorable local features or global features in different scales.
- During the inference phase of the UFCs, we used learnable parameters to select a better prediction result instead of simply adding the outputs of the two classifiers.
- We applied UFCs to recognize uterine fibroids using hysteroscopy images to assist doctors in diagnosis.

2.2 Outline

Section 2.1 introduces the research topic of UF, emphasizing their prevalence and associated complications concerning women's health. It further explores the potential utility of deep learning in the diagnosis of UF. The primary objective of this research is to develop a hybrid model that integrates CNN and Transformer for the detection of UF in hysteroscopic images.

In section 2.3, prior research on the application of deep learning for diagnosing UF is reviewed. Various studies, including the MBF-CDNN method for classifying UF-based ultrasound images and the use of the VGGNet-16 model for uterine lesion classification, are considered. Additionally, it investigates existing research on the fusion of CNN and Transformer network structures for diverse medical imaging tasks. However, the literature review reveals a dearth of studies focusing specifically on the recognition of UF using hysteroscopic images with a CNN and Transformer hybrid.

This section 2.4 offers a comprehensive description of the datasets utilized. The architecture of the proposed hybrid network is meticulously described. The metrics employed to assess the model's performance are delineated in this section.

The study's findings are presented in this section 2.5. The proposed model's performance is contrasted with extant architectures including the ConvNeXt network, Swin Transformer, and the original Conformer. The model proposed herein outperforms the others in terms of accuracy, sensitivity, specificity, F1-score, AUC, and precision.

This final section 2.6 engages in a discussion of the study's findings and their implications. The proposed CAD subsystem demonstrated exceptional performance in the identification of UF, offering higher accuracy, specificity, F1-score, AUC, and precision compared to other models. The balance between performance and computational efficiency achieved by the model is also highlighted. The potential of the model as a diagnostic tool in the medical field is underlined, and future plans for refining the model and validating its performance on larger and more diverse datasets are outlined.

2.3 Literature Review

Application of Deep Learning in Uterine Fibroids

Dilna et al. [79] proposed an MBF-CDNN method to classify UF-based ultrasound images. They used 259 ultrasound images and took 20% of them as the test set. The sensitivity of their test set is 94.44%, specificity 95 %, and accuracy 94.736%. They confirmed that their method had a better effect than the strong baseline CNN method.

Zhang et al. [80] used hysteroscopic images as the input of the VGGNet-16 model to classify uterine lesions. A total of 250 images were used as a test set to evaluate the performance of the model. The specificity of the model for UF was 80.0%. To segment UF in hysteroscopic images, Török et al. [81] applied manually annotated images and manually drawn bitmasks to train a fully convolutional neural network.

Combination of CNN and Transformer

Peng et al. [1] proposed to fuse the features obtained by the convolution operation with the features obtained by self-attention. In addition, some researchers have conducted research on the combination of CNN and Transformer network structures [82, 83]. The CNN-Transformer combined model was used for brain tumor segmentation on multi-modal MRI scans [84]. They proposed a fusion framework to segment CT images of human multiple organs. Xie et al. proposed a fusion framework to segment CT of human multiple organs [85]. Although deep learning has been widely used in the diagnosis of uterine diseases, relatively few studies were performed on the recognition of UF using hysteroscopic images. In addition, many researchers have explored the method of combining CNN and Transformer, but rarely used it in the medical field. Therefore, we implemented the combination of CNN and Transformer in recognizing UF.

2.4 Methods

2.4.1 Datasets

Our study utilized data collected between 2008 and 2019 from two prominent medical institutions: the Tongji Hospital of Huazhong University of Science and Technology (TJH) and the Maternal and Child Hospital of Hubei Province (MCH). From the hysteroscopic data provided by these hospitals, we curated a dataset comprising 5,868 images of UF from 273 patients, and 5,968 images of normal uteruses from 235 patients.

The dataset was then partitioned into training and test sets, with the division based on individual patients. This patient-based division ensures that the model is trained and tested on images from different subjects, thereby enhancing the generalizability and robustness of the model. Detailed information about the composition of the training and test sets can be found in Table 2.1.

In the process of preparing the images for the model, we implemented a cropping procedure to isolate the regions of interest and eliminate any extraneous non-lesion areas. This step ensures that the model focuses on the critical areas of each image, thereby improving its ability to detect and classify uterine fibroids. For the purposes of training, validation, and testing, all images were resized to a uniform dimension of 224×224 pixels. This standardization of image dimensions is a preprocessing step, ensuring that the model can process all images consistently and effectively.

Table 2.1: Details of the uterine fibroid dataset for classification

	Dataset		Training set		Test set	
	UF	Normal	UF	Normal	UF	Normal
Patients	273	235	240	199	33	36
Images	5,868	5,968	4,726	4,798	1,142	1,170

2.4.2 Network Structure

We constructed the CNN and transformer branches of our UF classification system based on Peng et al.'s work in [1]. The Conformer network is known for its ability to fuse local features and global representations under different resolutions interactively, and it adopts a concurrent structure to retain local features and global representations to the maximum extent.

The network structure of original Conformer as shown in Figure 2.1. This system is a hybrid, leveraging the strengths of both Convolutional Neural Networks and Transformers, and is designed to optimize the classification task. The Conformer network can be divided into two main branches: the CNN branch and the Transformer branch. The

CNN branch consists of four bottleneck stages, each designed to extract and process different levels of features from the input images. On the other hand, the Transformer branch is composed of four transformer blocks, each responsible for creating different feature representations. Figure 2.3 illustrates the proposed network structure. The preprocessed fibroids images are input to a common stem layer, which includes a 7×7 convolutional layer and a 3×3 pooling layer. This initial layer is crucial as it extracts edge and texture information from the images, which are fundamental features for the subsequent stages. Then, the extracted features are then fed into the CNN and Transformer branches. These branches work in tandem to process the features and generate different feature representations. To facilitate the exchange of heterogeneous information between the two branches and to address the misalignment between CNN feature maps and patch embeddings, a feature coupling unit is needed to handle the problem. This unit is placed between the outputs of each CNN stage and Transformer block. However, it's worth noting that the original feature coupling unit, in its current form, always adds the aligned feature information equally under different scales. This could potentially lead to scale variation issues under different bottleneck stages and transformer blocks. Therefore, further improvements to this unit are necessary to optimize the performance of our system.

In order to tackle the challenge of scale variation, we have integrated a weighting gate mechanism within our system. This mechanism, armed with a set of trainable parameters, is engineered to autonomously ascertain the fusion weight for each branch at various feature levels. This represents a substantial enhancement from the original methodology, where the outputs of the two branches were merely aggregated. The architecture of our enhanced Conformer, complete with learnable parameters, is depicted in Figure 2.2.

Our UF classification system has been realized using the PyTorch deep learning framework, a popular choice for its flexibility and efficiency. For optimization, we employ a Stochastic Gradient Descent (SGD) approach with an initial learning rate set at 0.01 and an L2 penalty of 5e-4 to prevent overfitting. To dynamically adjust the

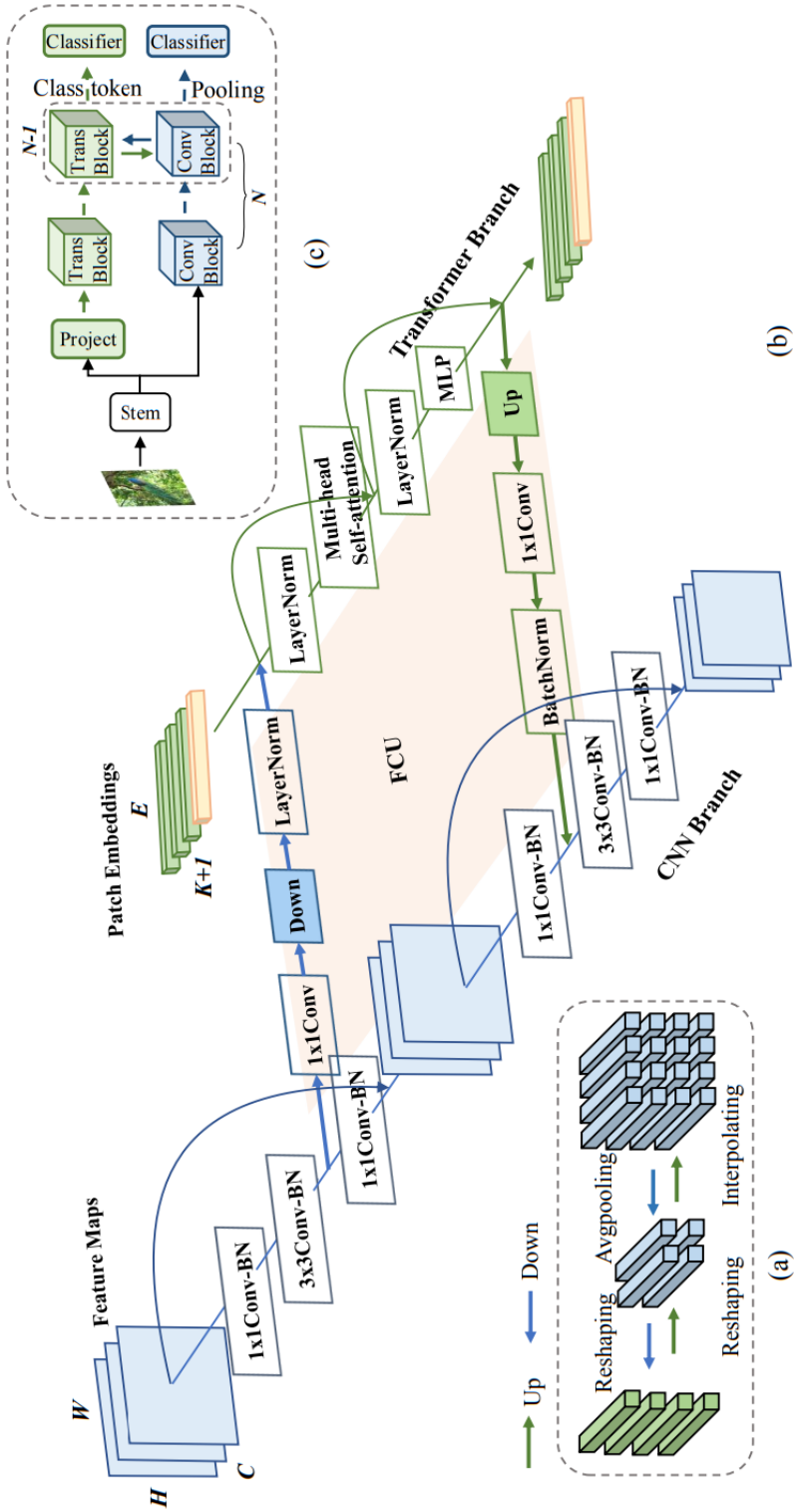


Figure 2.1: Network structure of the original Conformer proposed by Peng et al. [1]. (a) Spatial alignment of feature maps and patch embeddings through up-sampling and down-sampling techniques. (b) Detailed implementation considerations for the CNN block, transformer block, and Feature Coupling Unit (FCU). (c) An overview of the original Conformer structure.

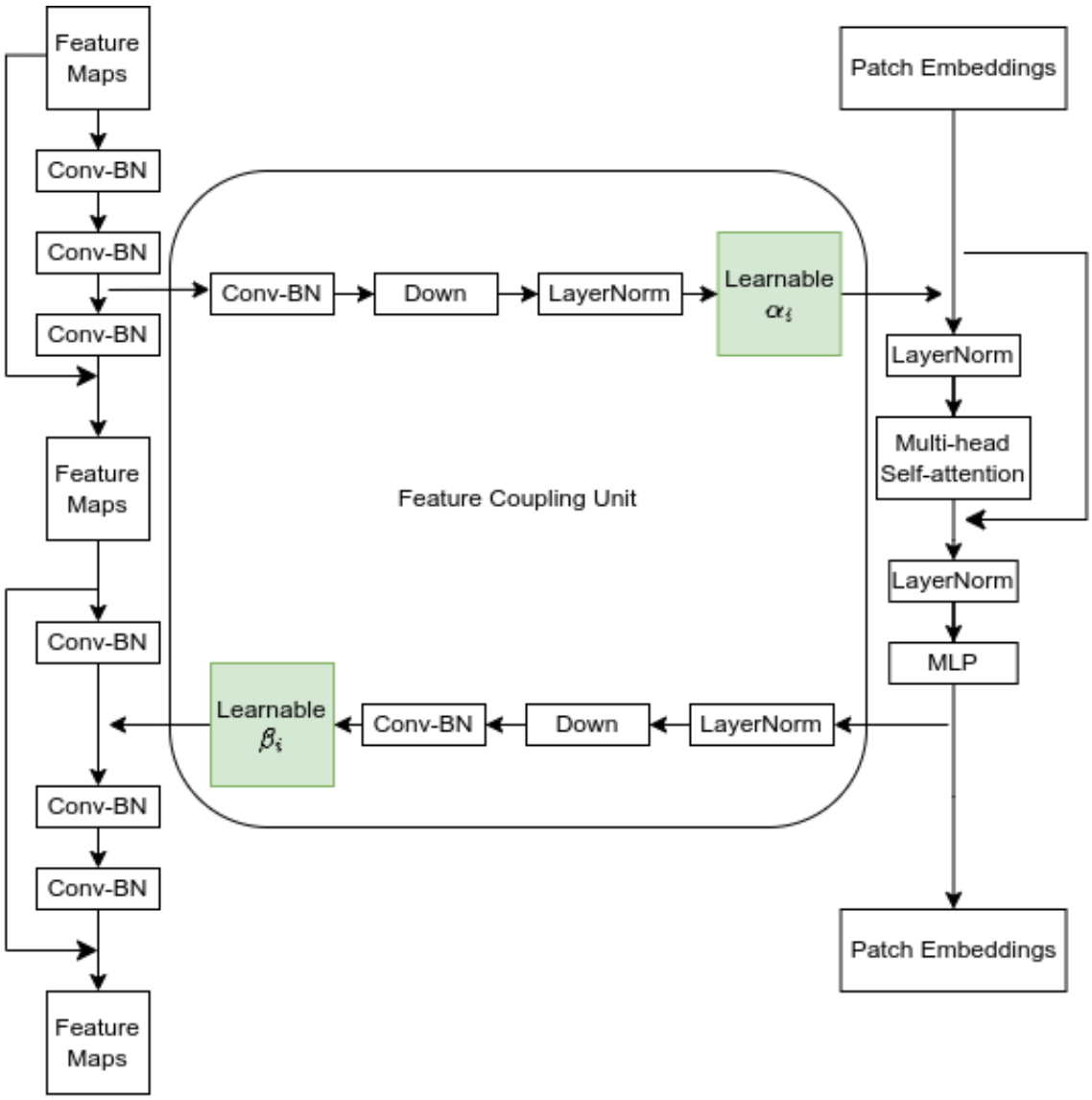


Figure 2.2: Network structure of our improved Conformer.

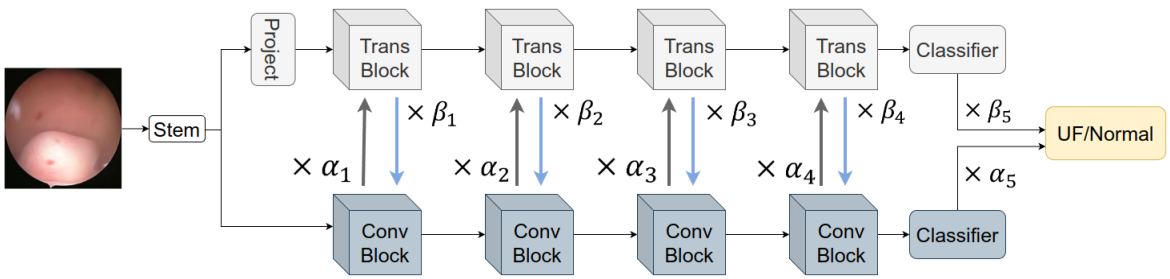


Figure 2.3: Network architecture of UFCs. UFCs is a hybrid network composed of a transformer branch and a CNN branch. α and β are a set of learnable parameters controlling the contribution of Trans and Conv Blocks, respectively.

Table 2.2: Diagnostic performance of UFCs and other deep learning models in uterine fibroids classification.

Model	Sensitivity	Specificity	Accuracy	F_1-score	AUC	Precision
ConvNeXt	0.8739	0.7274	0.7997	0.8117	0.8957	0.7578
Swin Transformer	0.8529	0.7137	0.7824	0.7948	0.9010	0.7441
Conformer	**0.9562**	0.8043	0.8793	0.8867	0.9646	0.8266
UFCs(Ours)	0.9421	**0.8376**	**0.8893**	**0.8936**	**0.9649**	**0.8499**

learning rate during the training process, we use a cosine annealing scheduler, a method known for its effectiveness in achieving better generalization accuracy. The training process is designed to run for 200 epochs, with a batch size of 32. The settings of these hyperparameters remain consistent throughout the training phase, ensuring stability and reproducibility of our results.

2.4.3 Evaluation Metrics

The performance metrics employed in our evaluation encompassed sensitivity, specificity, accuracy, F1-score, Area Under the Curve (AUC), and precision. These metrics provide a comprehensive understanding of the model's performance, from its ability to correctly identify positive cases (sensitivity) and negative cases (specificity) to its overall correctness (accuracy), the balance between precision and recall (F1-score), and overall performance (AUC). Additionally, we utilized a confusion matrix, which includes the counts of True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN).

$$Sensitivity = TP/(TP + FN) \tag{2.1}$$

$$Specificity = TN/(TN + FP) \tag{2.2}$$

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) \tag{2.3}$$

$$Precision(PPV) = TP/(TP + FP) \tag{2.4}$$

$$F_1 - score = 2 \times Sensitivity \times Specificity/(Sensitivity + Specificity) \tag{2.5}$$

2.5 Results

For the purpose of evaluating our model, we utilized a test set comprising 1142 uterine fibroid images sourced from 33 patients, along with 1170 uterine images obtained from 36 healthy subjects. We compare the performance of the proposed UFCs with the ConvNeXt network [86] represented by CNN architecture, the Swin Transformer [36] represented by transformer architecture, and the original hybrid network Conformer.

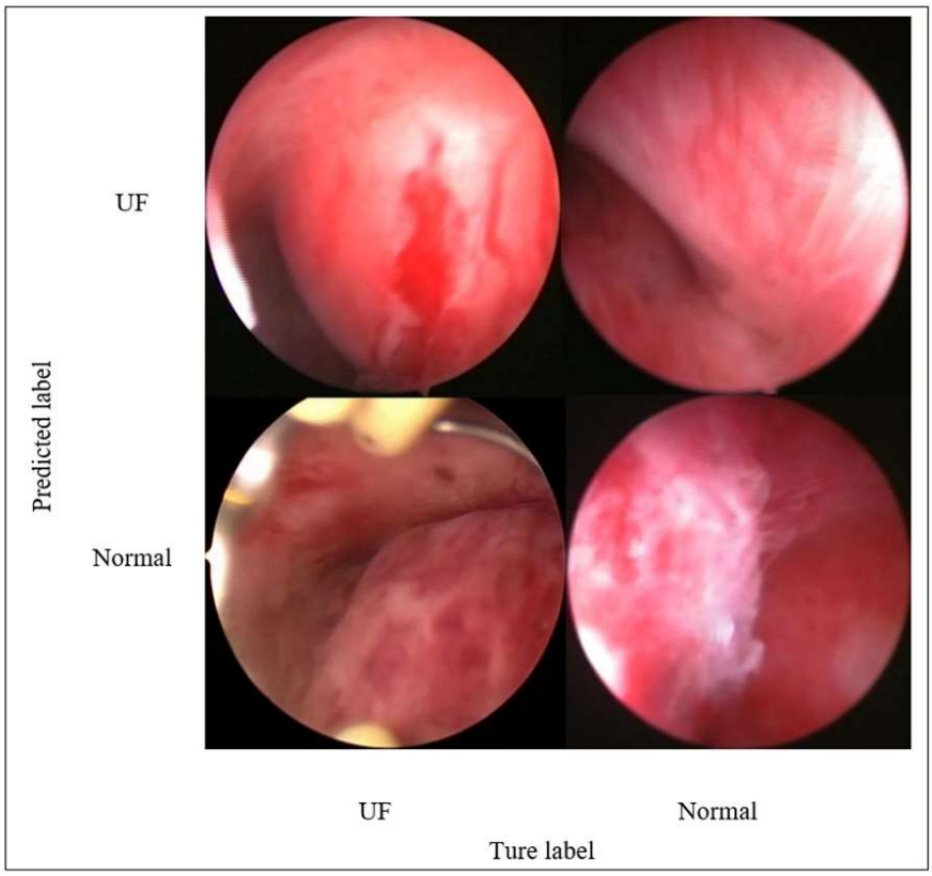


Figure 2.4: Example classification results output by UFCs. The horizontal axis is the true label of the model's output and the vertical axis is the predicted label by the model. UF and Normal indicate uterine fibroid and normal uterine images, respectively.

Table 2.2 lists the test results of different models. The test results demonstrate that a single ConvNeXt network or Swin Transformer network does not perform as effectively as hybrid networks. Especially in terms of sensitivity and accuracy, they performed far worse than hybrid networks. Table 2.3 lists that there are no obvious gaps between the parameter amounts (Params) and floating-point operations (FLOPs) among different models. Although the proposed method UFCs uses some learnable parameters than

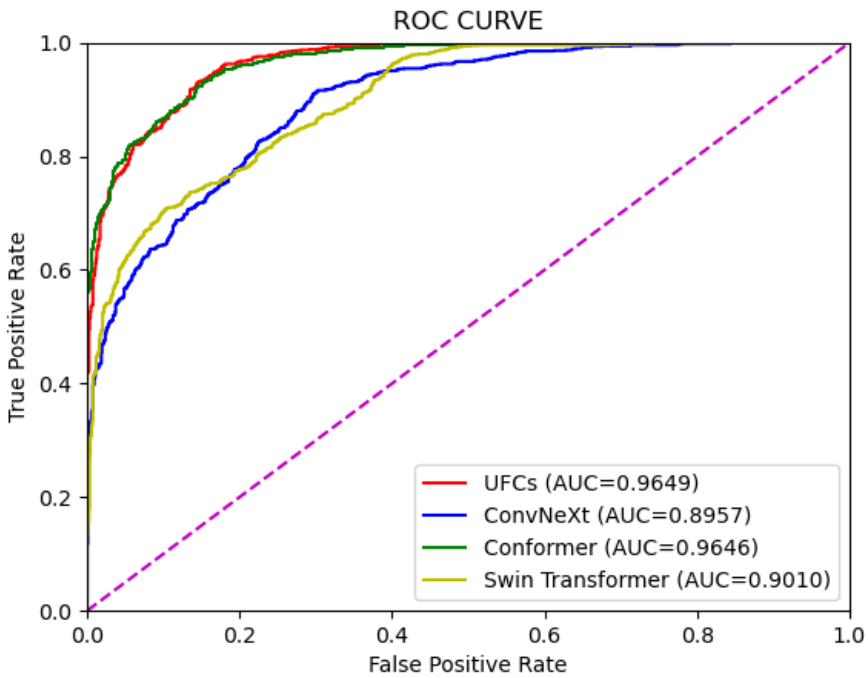


Figure 2.5: The comparison of receiver operating characteristic (ROC) curves.

the Conformer, the effects on the Params and FLOPs are negligible. Compared with the Conformer, the sensitivity of UFCs is slightly lower, and the performance of other metrics are superior to the Conformer former. On a test set of 2312 hysteroscopic images of UF, UFCs has a 0.9421 sensitivity, 0.8376 specificity, 0.8893 accuracy, 0.8936 F_1-score, 0.9649 AUC, and 0.8499 accuracy. Some examples of the classification results of UFCs on the test set are displayed in Figure 2.4.

Table 2.3: Comparison of different models.

Model	Params	FLOPs
ConvNeXt	29M	4.5G
Swin Transformer	29M	4.5G
Conformer	23.5M	4.9G
UFCs(Ours)	23.5M	4.9G

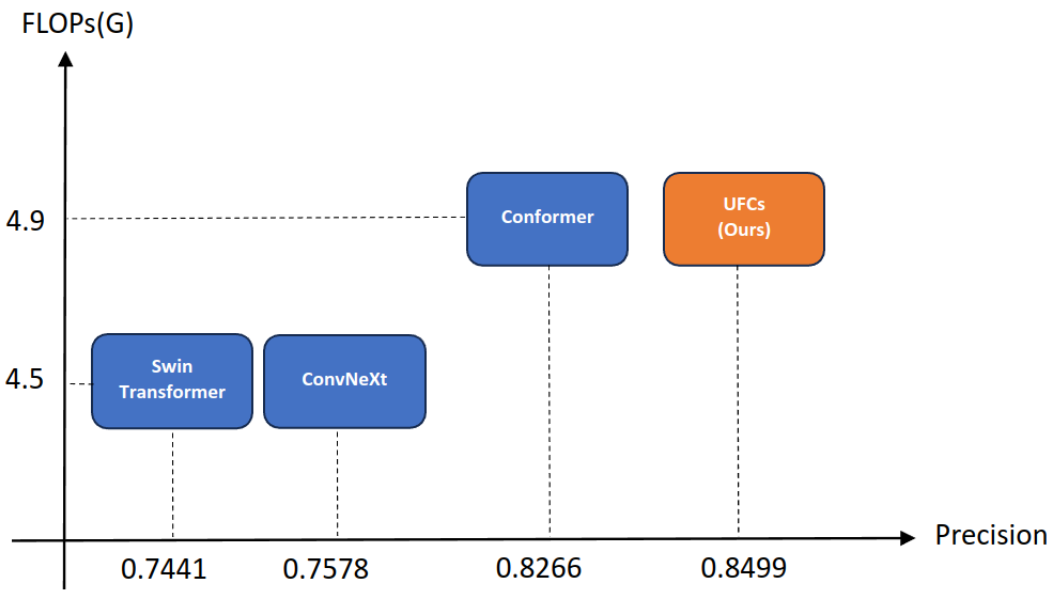


Figure 2.6: Our model is compared with ConvNeXt, Swin Transformer, and Conformer in terms of precision and FLOPs metrics.

2.6 Discussion

In this study, we have introduced a CAD subsystem, an amalgamation of Convolutional Neural Networks and Transformer models, specifically tailored for the identification of uterine fibroids. This approach represents a significant advancement in the field of medical image classification, particularly in the context of uterine fibroid recognition.

Upon comparison with existing models, our proposed CAD system exhibited superior performance, achieving an accuracy of 0.8893. This is a noteworthy improvement over the ConvNeXt model, which achieved an accuracy of 0.7997, and the Swin Transformer model, which achieved an accuracy of 0.7824. Even when compared to the original Conformer model, our system demonstrated a higher accuracy, with the Conformer model achieving an accuracy of 0.8793. Furthermore, our model also outperformed the other models in terms of specificity, F1-score, Area Under the Curve (AUC), and precision, further underscoring its superior diagnostic performance. Specifically, our model achieved an F1-score of 0.8936, an AUC of 0.9649, and a precision of 0.8499, all of which were the highest among the compared models. In terms of computational efficiency, our model matches the original Conformer model with 23.5M parameters and 4.9G FLOPs, demonstrating that our system maintains a balance between performance

and computational efficiency.

The collective findings from our study underscore the effectiveness of our proposed CAD system in the recognition of uterine fibroids. Our model's superior performance, both in terms of diagnostic accuracy and computational efficiency, sets it apart from existing models. It not only outperforms them in key metrics but also demonstrates a balance between performance and computational demands, a critical factor in real-world applications. The high diagnostic accuracy of our model, as evidenced by its performance on various metrics, makes it a promising tool in the medical field. Its ability to accurately identify uterine fibroids can aid clinicians in making more informed diagnoses, potentially leading to more effective treatment plans and improved patient outcomes. Our future work will focus on further refining the model and expanding its capabilities. We plan to validate its performance on larger and more diverse datasets, which will provide a more robust test of its generalizability and effectiveness. We also aim to explore ways to further optimize the model, such as fine-tuning the weighting gate mechanism or exploring different fusion strategies.

2.7 Conclusion

In conclusion, this study has successfully demonstrated the efficacy of a CAD subsystem, a hybrid of Convolutional Neural Networks and Transformer models, in the detection of uterine fibroids. Our CAD model has outperformed several existing models, including ConvNeXt, Swin Transformer, and even the original Conformer model, in terms of diagnostic accuracy, specificity, F1-score, AUC, and precision, highlighting its superior diagnostic performance. This model's ability to accurately recognize uterine fibroids has the potential to assist healthcare professionals in making more informed diagnoses. Future work will aim to enhance the model further, test its generalizability and effectiveness on larger and diverse datasets, and explore optimization strategies, thereby solidifying its potential as a powerful tool in the realm of medical diagnostics.

Chapter 3

Computer-Aided Detection of Endometrial Polyps Using Deep Learning

3.1 Introduction

EP are defined as overgrowths of endometrial glands, stroma, and blood vessels from the lining of the uterus. Women of all ages may suffer from this disorder. The peak incidence age is 40–49 years [7,8]. The main symptoms include abnormal uterine bleeding, pelvic pain, and infertility. The malignant transformation rate of EP is in the range of 0–13% [9,10].

Hysteroscopic endometrial polypectomy is the standard treatment for patients with symptoms and high risks [87,88]. However, the complications of this approach—such as intraoperative bleeding, uterine perforation, peripheral organ damage, water intoxication, and intrauterine adhesions—should be of great concern [19]. Water intoxication can induce systemic toxicity and even death, caused by the fluid loading dury the surgery. Experienced hysteroscopists play an important role in reducing these intra-

operative and postoperative complications and can make preliminary predictions about the extent of the disease. Therefore, a long period of experience is required for effective hysteroscopy, while junior hysteroscopists may extend a procedure unnecessarily or miss polyps. Therefore, a tool that can assist in diagnosis is urgently needed to fill this gap.

Deep learning has powerful capabilities to learn and recognize patterns in data; therefore, it has been widely implemented in medical image analysis in recent years. Previous studies have been performed to optimize the efficiency of analysis, diagnosis, and treatment strategies using deep learning. Ramamurthy et al. proposed an Effimix model, combining squeeze and excitation layers, along with self-normalization activation layers, with a backbone of EfficientNet B0. Their model achieved a high accuracy of 97.99% in the classification of gastrointestinal diseases [89]. Muruganantham et al. combined ResNet-50 and self-attention mechanisms to localize gastrointestinal lesions in wireless capsule endoscopy images, achieving a classification accuracy of 95.1% and 94.7% on two publicly available datasets—the bleeding dataset and the Kvasir-Capsule dataset, respectively [90]. Object detection algorithms based on deep learning provide the location and classification of lesions in medical images as advantageous tools for aiding in clinical diagnosis. Jha et al. proposed ColonSegNet to detect, localize, and segment polyps in colonoscopy images, with a better trade-off between an average precision of 80.0% and mean IoU of 0.810, and a maximum speed of 180 frames per second for the detection and localization task [91]. More studies have demonstrated the usefulness of object detection algorithms for lesion detection in endoscopic images [92–94]. Some researchers have also applied deep learning to the identification of endometrial lesions. Hodneland et al. used a three-dimensional CNN to automate the segmentation of endometrial cancer in magnetic resonance images (MRIs) [95]. They claimed that the CNN-based segmentation accuracy and tumor volume estimation were equivalent to the results achieved by radiologists' manual segmentation. Kurata et al. segmented uterine endometrial cancer via multi-sequence MRI with a CNN [96]. Zhang et al. used a VGGNet-16 model to classify endometrial lesions from hysteroscopic images [80].

Takahashi et al. proposed a system based on deep learning of hysteroscopy images to localize endometrial cancer lesions automatically [97].

To date, endometrial lesion evaluation based on CAD has mainly focused on object classification or segmentation using computerized tomography images, MRI, and ultrasound images. The application of deep learning is rarely employed in the object detection of endometrial polyps using hysteroscopy. In this study, we investigate the CAD subsystem using deep neural networks for the identification and location of EP. This would assist doctors in detecting lesions, lessen their workload, improve the accuracy of diagnosis and treatment, and reduce the risk of endometrial cancer. Our contributions are summarized as follows:

- We proposed group normalization in the deep learning model YOLOX to improve the performance of real-time detection of endometrial polyps from hysteroscopic images.

- A video adjacent-frame association algorithm was applied in the post-processing stage. The algorithm effectively solved the problem of the original YOLOX, i.e., unstable polyp detection.

- We present the first application based on deep learning to detect endometrial polyps from hysteroscopic images.

3.2 Outline

The section 3.1 highlights the potential of deep learning in diagnosing EP, a prevalent gynecological disorder. Despite the efficacy of hysteroscopic endometrial polypectomy as a standard treatment, the associated risks necessitate a more reliable diagnostic tool. Leveraging the proven efficiency of deep learning in medical image analysis, this study introduces a novel CAD subsystem for identifying and locating EP. This subsystem enhances the YOLOX deep learning model through group normalization and

employs a video adjacent-frame association algorithm in post-processing to ensure accurate and consistent detection of EP from hysteroscopic images.

Section 3.3 outlines the research methodology involving data collection from two hospitals and the utilization of YOLOX using group normalization to allow smaller batch sizes. Post-processing employs a perceptual hash algorithm to enhance detection stability. A comprehensive description has been provided for the definition of evaluation metrics at the video and image levels.

In this section 3.4, the modified YOLOX model was assessed through ablation experiments and showed efficiency in the detection of EP. The modifications allowed for reduced training time and fewer epochs. Furthermore, it was demonstrated that the Group Normalization and VAFA algorithm increased model efficiency. Ablation studies also showed improved performance metrics for the modified model across two different test sets. The study further showed the model's capacity for real-time endometrial polyp detection, indicating its suitability for practical medical applications.

This section 3.5 discusses the application of the enhanced YOLOX model for detecting EP in hysteroscopic videos. The proposed model exhibited superior sensitivity and real-time performance compared to the original YOLOX model. Its generalization capabilities, largely due to the use of the GN method and the VAFA algorithm, proved superior, particularly for the external test dataset from the TJH. The improved model also outperformed EfficientDet, another highly efficient model. Limitations of the model include poor performance with partially occluded images and misdiagnosis of a large floating endometrium. Future work includes integrating the algorithm into existing clinical systems and optimizing it for mobile or web-based platforms.

3.3 Methods

3.3.1 Datasets

The training and test datasets were collected from a consecutive series of patients at the Maternal and Child Hospital (MCH) in Hubei Province during 2008–2019 and Tongji Hospital (TJH) at Huazhong University of Science and Technology during 2018-2020.

The training set included videos from 323 cases diagnosed with polyps in the MCH. A total of 7313 and 4526 images with and without polyps, respectively, were extracted from these cases for training. Our test sets comprised cases from two hospitals (MCH and TJH). The MCH test set served as the internal test data and comprised videos from 48 and 183 cases with and without polyps, respectively. For the external test data, the TJH test set comprised videos from 150 and 50 cases with and without polyps, respectively. In addition, we employed five videos with polyps from the TJH to evaluate the real-time detection performance of the proposed model. Figure 3.1 displays examples of the hysteroscopic images used in this study.

Figure 3.2 depicts the general processing flow of the proposed method. Videos from the two test datasets were utilized to evaluate the accuracy, specificity, and sensitivity of the model.

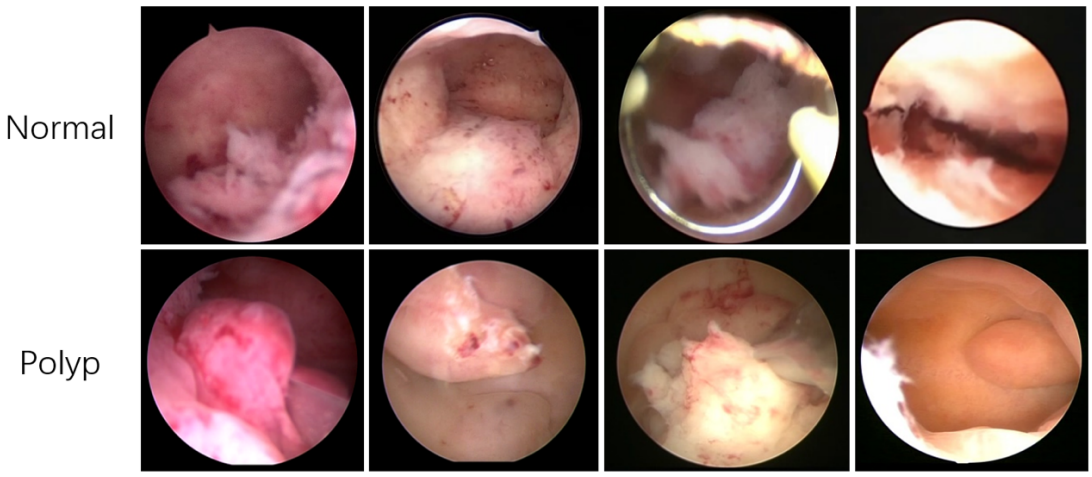


Figure 3.1: Examples of the hysteroscopic images used in this study.

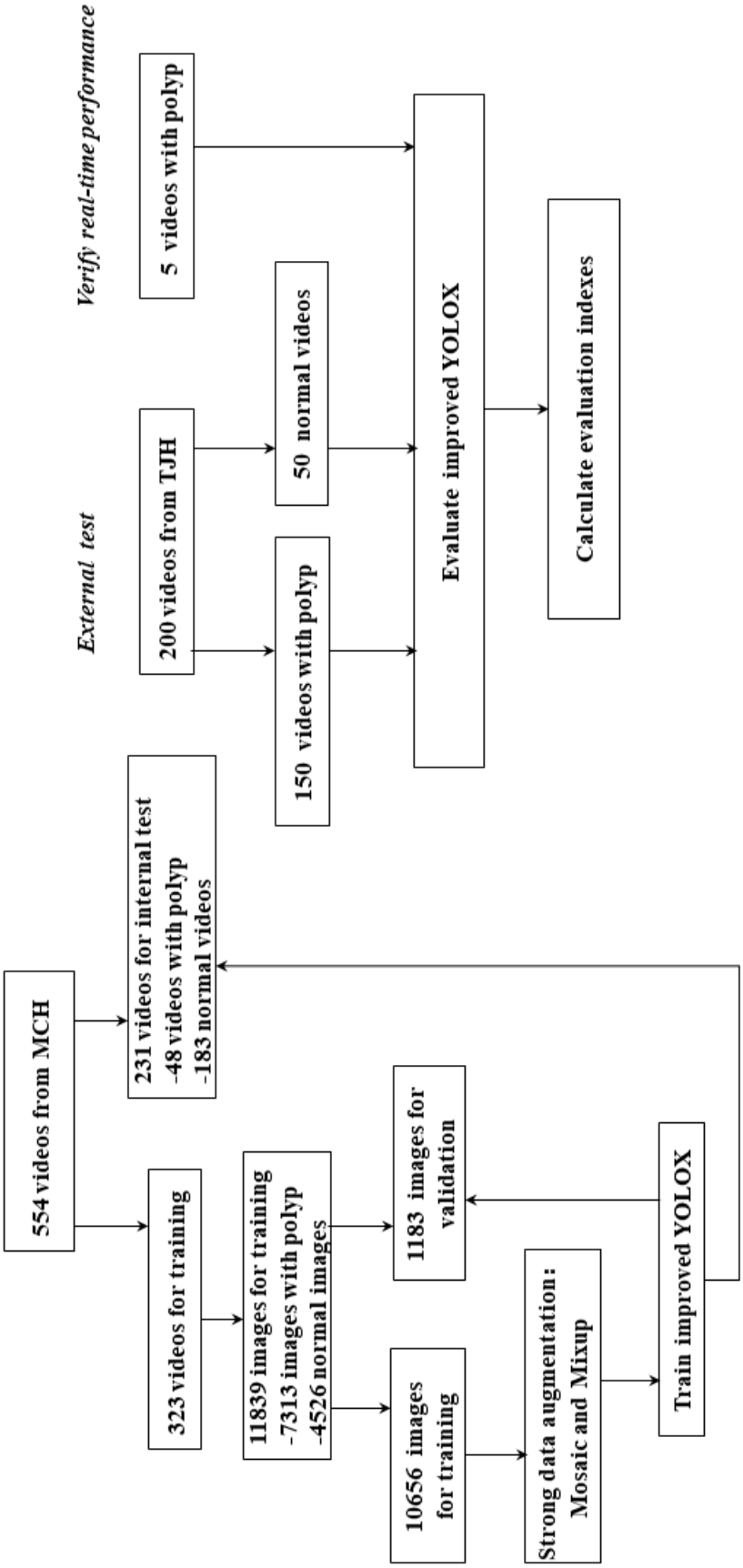


Figure 3.2: Development and validation flowchart. MCH, Maternal and Child Hospital; TJH, Tongji Hospital.

3.3.2 Data Preprocessing

Two gynecologists (W.W. and X.D.) annotated the training set manually, selecting normal images and images containing at least one polyp. To enhance the generalization and robustness of the deep learning model, mosaic and mixup augmentation methods were used in data preprocessing [98].

Mosaic augmentation involves the following steps:

(1) randomly select one set of coordinates from the coordinates of the center points of four images to be included in the output image;

(2) randomly choose the indices of the other three images and read their corresponding labels;

(3) resize each image to 640 $\times$ 640 pixels while preserving its aspect ratio;

(4) based on the rules of up, down, left, and right, calculate the position where each image should be placed in the output image;

(5) crop the output image.

Mixup augmentation is implemented as follows:

(1) use random jitter augmentation;

(2) randomly apply flip augmentation;

(3) mix the original image and the processed image using a ratio of 1:1.

The data augmentation implementation process is shown in Figure 3.3.

3.3.3 Improved YOLOX

YOLOX [98] represents a cutting-edge development in the field of real-time object detection, leveraging the power of CNN to identify and classify objects within images and videos. YOLOX has been recognized for its exceptional performance and processing speed, making it a valuable tool in real-time object detection applications. The

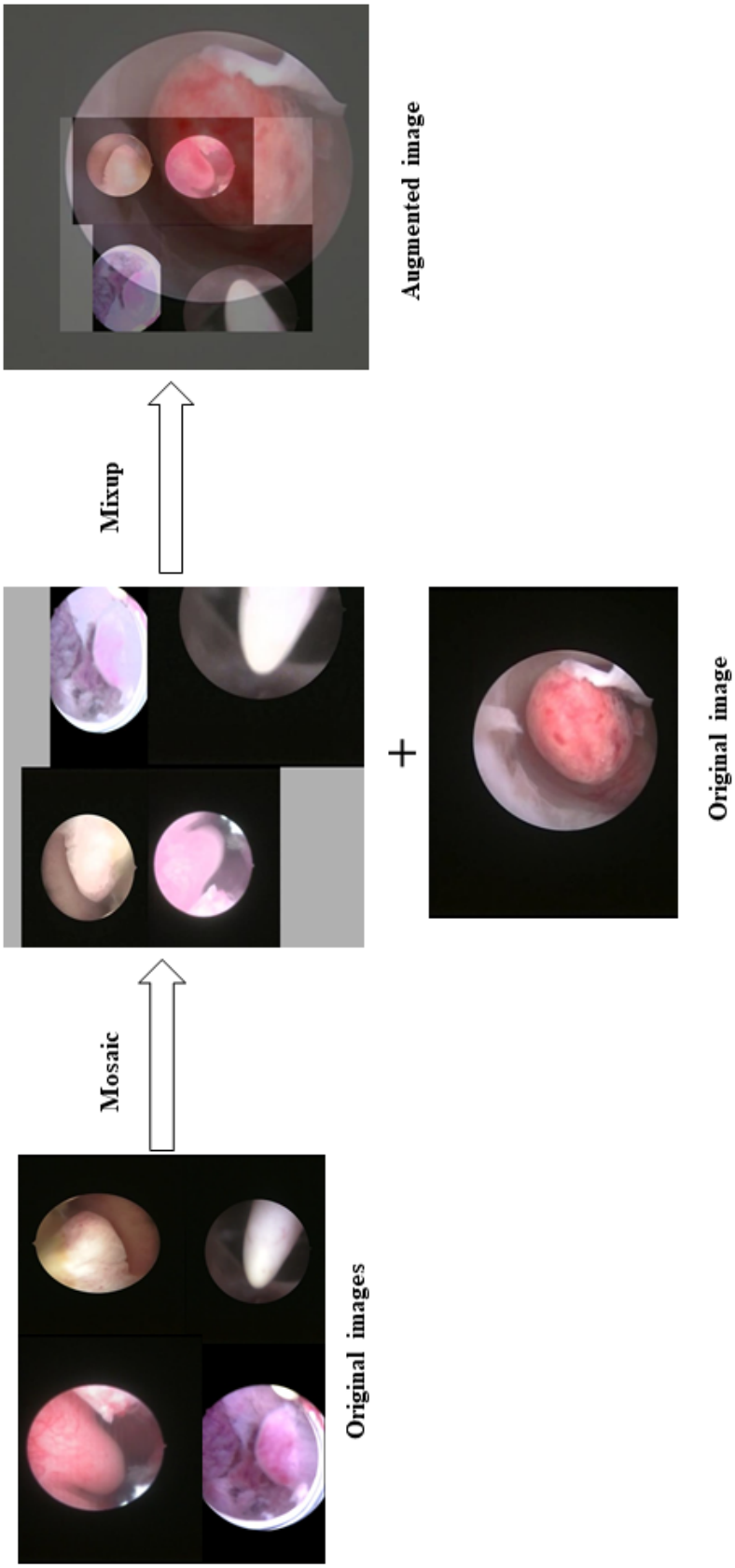


Figure 3.3: Data augmentation implementation process.

YOLOX model is an evolution of the original YOLO (You Only Look Once) models, introducing significant improvements that enhance its performance. One of the key enhancements is the adoption of an anchor-free design, which eliminates the need for predefined anchor boxes and allows the model to generate predictions directly from the feature maps. This design simplifies the detection process and improves the model's flexibility in handling objects of various sizes. In addition, YOLOX introduces a decoupled head structure, which separates the prediction of object classes and bounding boxes. This decoupling allows the model to optimize each task independently, leading to more accurate and reliable predictions. Furthermore, YOLOX employs robust data augmentation techniques, enhancing the model's ability to generalize and perform well on diverse and unseen data.

In the context of this study, we propose a model based on YOLOX for hysteroscopic detection. This application represents a novel use of YOLOX, extending its capabilities into the realm of medical imaging. The proposed model has the potential to significantly enhance the diagnostic capabilities of clinicians during hysteroscopic surgery. By providing real-time, accurate detection of abnormalities, the proposed method can aid clinicians in making more informed decisions during surgery, potentially improving patient outcomes. As we continue to refine and validate this model, we anticipate that it will become an invaluable tool in the field of hysteroscopic surgery.

Group Normalization

The original YOLOX used batch normalization (BN) to improve accuracy [99]. However, BN requires a sufficiently large batch size to work effectively, because normalization is performed over the entire batch of input data. A small batch may lead to an inaccurate estimation of batch statistics and, therefore, increase errors. However, increasing the batch size may reduce the training effect and stability. As the batch size increases, more memory is required to store and process the data. If the memory capacity is insufficient, memory errors and crashes may occur. For example, because all hysteroscopy images are resized to 640 $\times$ 640 pixels in this study, a large batch size

is required for effective training, which may lead to insufficient memory. As a result, batch size and memory capacity should be balanced to achieve optimal training performance. To meet this challenge, a group normalization (GN) method was adopted to use smaller batch sizes without sacrificing performance [100]. The main difference between GN and normal normalization methods lies in using Si at each pixel point, where Si is the set of pixels by which the mean and standard deviation are calculated for normalization. A four-dimensional vector $i = (i_N, i_C, i_H, i_W)$ is used to index the features in the order of (N, C, H, W), where N represents the batch axis, C the channel axis, and H and W the spatial axes of height and width, respectively. When using GN, the value of S_i is defined as follows:

$$S_i = \left\{ k | k_N = i_N, \lfloor \frac{k_C}{C/G} \rfloor = \lfloor \frac{i_C}{C/G} \rfloor \right\} \tag{3.1}$$

where G is the number of groups (set to 32 by default), C/G is the number of channels per group, "$k_N = i_N$" refers to pixels within the same batch, and "$\lfloor \frac{k_C}{C/G} \rfloor = \lfloor \frac{i_C}{C/G} \rfloor$" refers to pixels with the same channel index and within the same group.

The GN method divides the channels into groups and calculates the mean and variance of each group for normalization. The main advantage of GN is that it normalizes the activations within each group rather than over batch sizes. Using GN, effective training can be achieved with relatively smaller batch sizes to improve the efficiency and effectiveness of YOLOX.

The network structure of the improved YOLOX is shown in Figure 3.4. The model network architecture consists of three main components: the backbone, neck, and decoupled head. The backbone network is responsible for extracting image features and comprises cross-stage partial connections (CSP), spatially separable convolution (SSP), bottleneck modules, and focus modules [98]. The CGL module is the smallest component in the network architecture of YOLOX. It is composed of convolution operations, GN, and SiLU activation functions. The detailed structure of the SPP and focus modules is shown in Figure 3.5 in the Appendix. CSP modules enhance fea-

ture complexity and diversity by sharing information across multiple layers through cross-stage local connections. SSP modules employ spatially separable convolutions to reduce model parameters and improve computational efficiency. Bottleneck modules increase feature expressiveness and reduce computational costs by employing dimensionality reduction and expansion techniques. The focus module—a specialized convolution module—decomposes input feature maps, thereby reducing convolutional computation while improving the detector's sensitivity to small objects.

The neck component is designed to generate a top-down pathway, which starts with the deepest feature map from the backbone network and progressively upsamples the feature map to produce higher-resolution feature maps. Simultaneously, a lateral connection pathway merges the upsampled features with corresponding features from the bottom-up pathway at each level. This process ensures that the multiscale feature maps retain high-level semantic information while incorporating finer details from lower-level features. Multiscale feature maps are utilized to further enhance the model's performance. The decoupled head module serves as the core of the detector, responsible for the classification and regression of features. This module adopts a decoupled head and anchor-free design, which not only reduces the number of parameters and increases the detection speed but also improves the detection accuracy. Moreover, an advanced label-assignment strategy is employed to address label imbalance and reduce the detector's propensity for errors, thereby elevating the model's performance.

VAFA Algorithm

In the post-processing stage, we used a perceptual hash algorithm (PHA) to address the issue of unstable object detection boxes. The PHA calculates the similarity of adjacent frames and uses this information to stabilize the detection boxes in similar frames. This helps to improve the accuracy and consistency of our object detection results. The PHA uses a feature vector called a "fingerprint" to represent an image. It then compares the fingerprints of different images to determine their similarity. The smaller the difference between the fingerprints, the more similar the images. After extensive test-

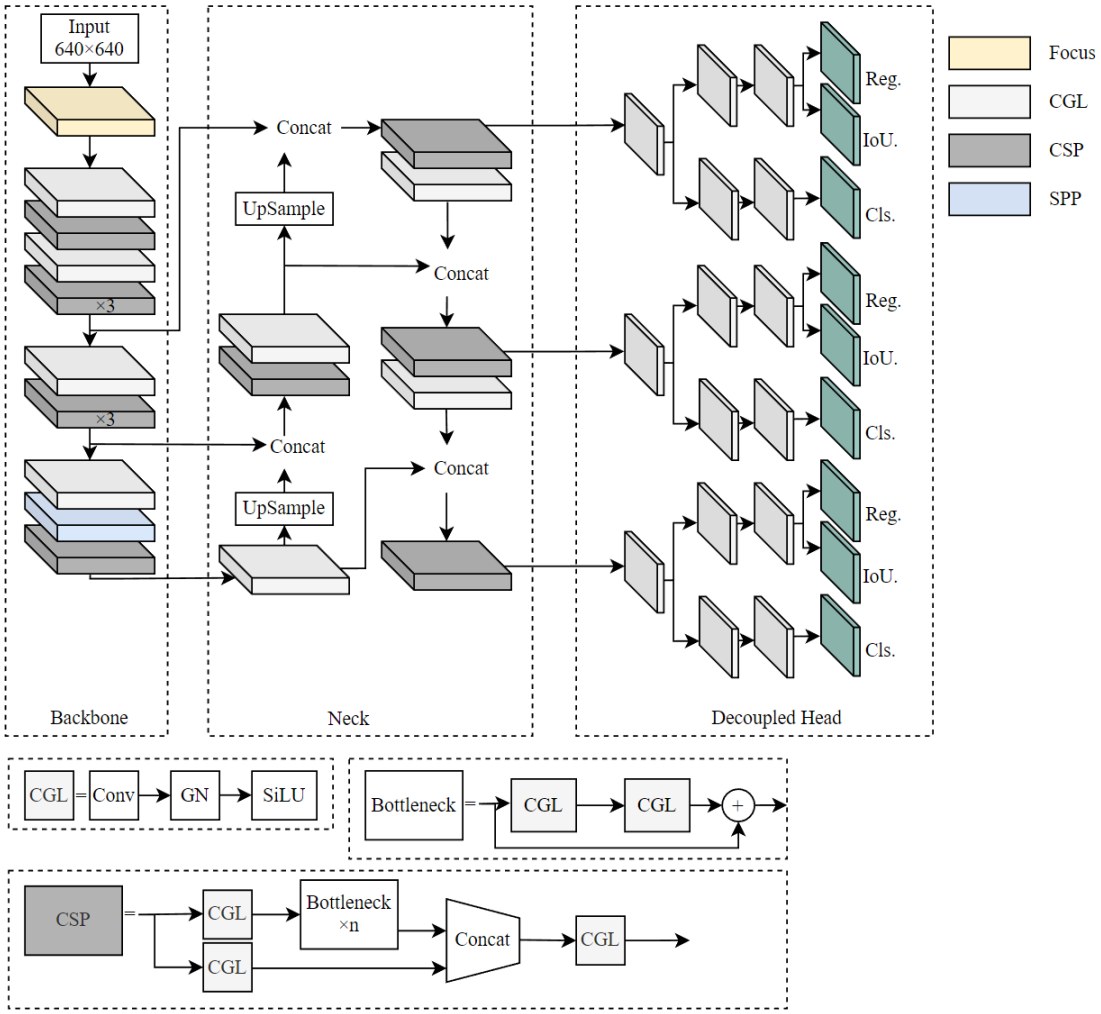


Figure 3.4: The network structure of the improved YOLOX. Modules are differentiated by color, as shown in the legend. The training images are resized to 640×640 pixels as the inputs of the model. The symbol "×3" means that there are three bottleneck modules.

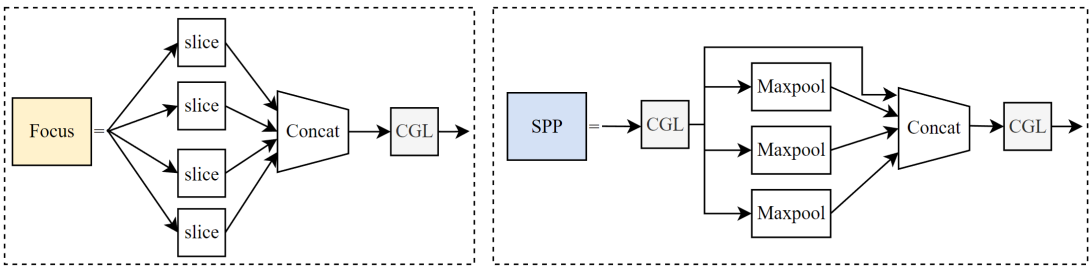


Figure 3.5: The detailed structure of the focus and SPP modules.

ing, we found, through trial and error, that when the Hamming distance between two images was ≥9, the images would no longer be considered similar. Using this threshold enabled us to identify and fix unstable detection boxes in our object detection results. This algorithm is defined as a video adjacent-frame association (VAFA) algorithm. Its implementation is as follows:

1. When five or more adjacent frames are detected as containing a "polyp," the similarity between the current frame and the next frame is calculated.

2. If the similarity is <9, the object detection box of the current frame is assigned to the next frame until the similarity between frames is no longer <9.

This process helps to stabilize the detection boxes and improve the accuracy of our object detection results, as shown in Figure 3.6.

3.3.4 Model Training

The loss function L for the proposed model comprises three components: bounding box regression loss L_{reg}, classification loss L_{cls}, and object loss L_{obj}.

$$L = \frac{reg_{weight} * 1}{N_pos} L_{reg} + \frac{1}{N_{pos}} L_{cls} + \frac{1}{N_{pos}} L_{obj} \tag{3.2}$$

where N_{pos} is the number of positive labels in each training batch, and reg_{weight} is a balancing coefficient. The classification loss is used to classify an object to one of the predefined classes, and the bounding box regression loss refines the bounding box around an object to make its detected location more accurate. The object loss L predicts the probability of an object being present in an image. Its three components are combined to optimize the model and improve its performance in detecting objects in images. To enhance the regression loss with the other components of the loss function in the YOLOX model, reg_{weight} was set to 5.0 by trial and error. This ensures that the model can refine the bounding boxes around detected objects accurately and improve its performance in object detection.

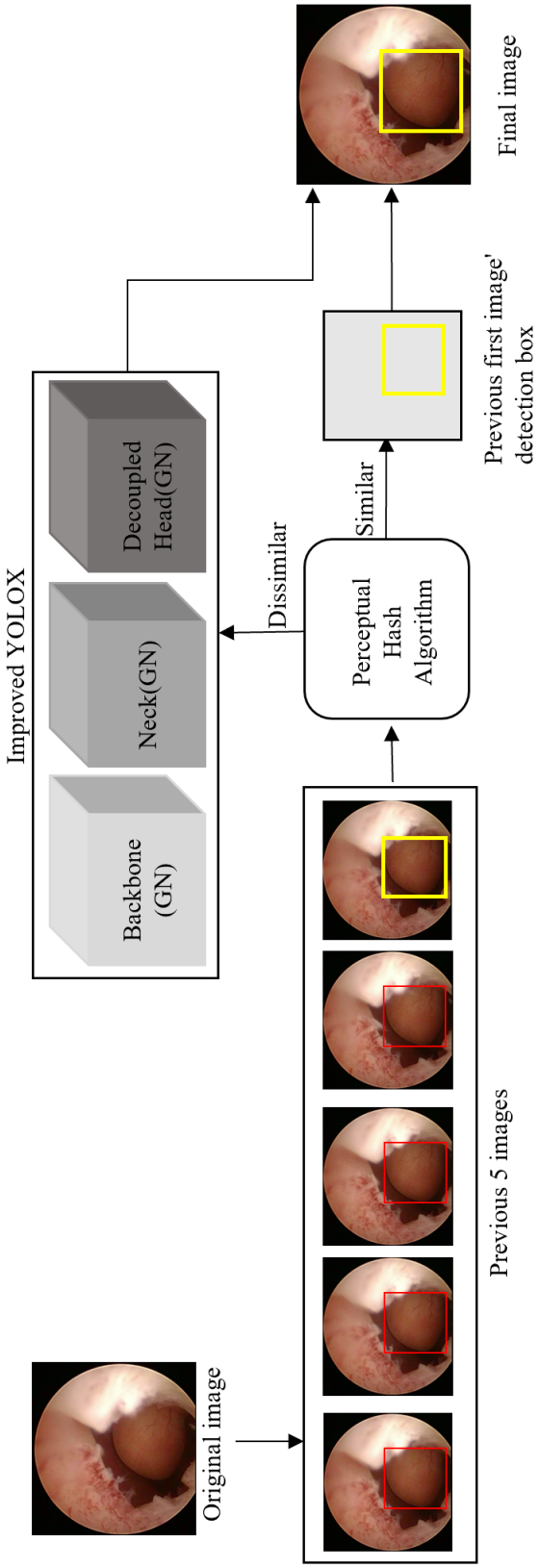


Figure 3.6: The detailed structure of the focus and SPP modules.

During the training phase, the MCH training data were randomly divided into training and validation sets using a 9:1 ratio. YOLOX-S was used as a baseline model for comparing performance. The input size for the neural network was 640×640 pixels, the input images were augmented automatically via the methods introduced in Section Data Preprocessing, and augmentation was turned off for the final 15 epochs. We specified the iteration rate as one per 10 epochs except for the last 15 epochs, when validating the validation set and calculating the mean average precision (mAP). In addition, validation was performed after each iteration of the last 15 epochs. The mAP was calculated by comparing a ground-truth bounding box with the detected box. The higher the mAP score, the more accurate the detection result. The weight values of the proposed model were saved, with the highest mAP value as the best checkpoint.

The models were developed using a Python 3.9 and PyTorch 1.12.1 framework and trained on a workstation equipped with an AMD Ryzen 7 3700X 8-Core processor CPU, 32.0 GB RAM, and an NVIDIA GeForce RTX 3060. The training batch size was set to eight, the number of epochs was 300, and half-precision training was enabled. These training hyperparameters remained unchanged during training and were kept constant when training YOLOX models with GN and the VAFA algorithm. Figure 3.7 illustrates the step-wise loss for both the original YOLOX model and the improved model.

3.3.5 Evaluation

We employed sensitivity, specificity, accuracy, precision, and F_1-Score to evaluate the performance of the proposed model. improved the YOLOX model to achieve higher sensitivity in the detection of endometrial polyps. The VAFA algorithm was also proposed in a post- processing stage to improve the stability of the detection process and enhance its convenience of use by hysteroscopists. The improved model and method showed significant generalizability and stable capacity and could achieve high sensitivity in the detection of endometrial polyps. Future works will mainly be based on the practical application of this study. Firstly, we expect to explore the feasibility of integrating our algorithm into existing clinical hysteroscopy systems to improve the ef-

ficiency and accuracy of endometrial polyp detection. Secondly, we want to optimize the proposed model by combining hysteroscopic videos and medical information for use on mobile devices or web-based platforms. This would facilitate the usage of the proposed method in remote healthcare services or telemedicine systems. Overall, the improved model may be useful in reducing the missed diagnosis rate in clinical hysteroscopic surgery and improving the detection sensitivity of endometrial polyps.

The evaluation metrics employed in this chapter align with the computational method described in Section 2.4.3. We evaluated the model at both video and image levels. To calculate evaluation metrics at the image level, the videos were converted to frames. If the bounding box output by the model overlapped with the location of the polyp in a frame and no bounding box appeared in non-polyp areas, the frame was recorded as a true positive (TP). If a polyp was detected in the wrong position, it was counted as a false positive (FP). If a frame did not actually contain a polyp and was detected as not having any polyps, it was recorded as a true negative (TN). Otherwise, it was considered a false negative (FN). At the video level, if the number of correctly detected frames (i.e., TP or TN) exceeded half of the total number of frames in the video, the video was considered TP or TN. Otherwise, the video was considered FP or FN.

3.4 Results

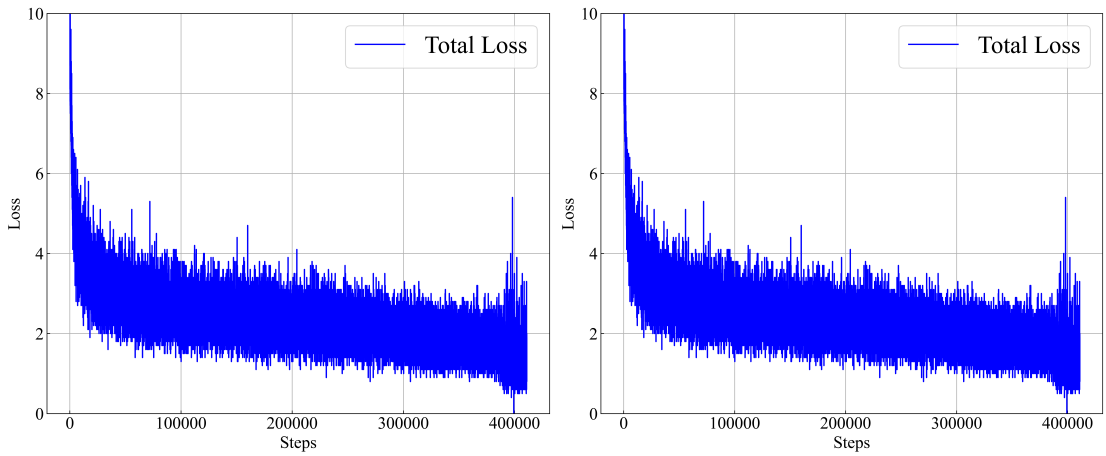
To assess the efficacy of the modifications to the YOLOX model, we applied it in ablation experiments and calculated the evaluation metrics described in Section Evaluation. The models were tested using the checkpoint with the highest mAP. The original model and the model using GN achieved the best checkpoint values after 286 and 260 epochs, respectively. The model using both GN and the VAFA algorithm reached the best checkpoint after 260 epochs, given that VAFA is a post-processing step and does not affect the training process of the model. The results indicated that the improved YOLOX model was able to fit the training data more efficiently, requiring fewer epochs and less time in training, as compared with the original model. This suggests that the

modifications were effective in reducing the training time.

To validate the efficacy of GN and VAFA, we conducted ablation studies. The performance of the four models, as shown in Table 3.1, was evaluated with the MCH and TJH test sets, with a confidence level of 0.4. The images were resized to 640×640 pixels for the evaluation. Both the image-level and video-level performances of the models were assessed. Each case in the test set was represented by only a single video.

Table 3.1 shows that the (per-lesion) sensitivity, specificity, accuracy, precision, and F1 for the YOLOX+GN+VAFA algorithm with the MCH test set were 100%, 88.52%, 90.91%, 69.57%, and 93.91%, respectively. With the TJH test set, the equivalent results for the YOLOX+GN+VAFA algorithm were 92.0%, 76.0%, 88.0%, 92.0%, and 83.24%, respectively. As shown in Table 3.2, the YOLOX+GN+VAFA algorithm had a per-image sensitivity of 98.73%, compared with the original YOLOX model's 96.01%, for the MCH test set. The original YOLOX model's per-image sensitivity with the TJH test set was 82.07%, whereas that for the YOLOX+GN+VAFA algorithm was 92.92%.

The proposed model is therefore capable of real-time endometrial polyp detection, with a video processing speed of 63 FPS using a NVIDIA GeForce RTX 3060 GPU. This makes it a suitable solution for practical medical applications.



(a) (b)

Figure 3.7: Stepwise loss curves. (a) The loss curve of the improved YOLOX model. (b) The loss curve of the original YOLOX model.

Table 3.1: Evaluation results of ablation experiments using the MCH and TJH test sets.

Model	Sensitivity(%)	Specificity(%)	Accuracy(%)	Precision(%)	F_1-score(%)
MCH test set					
YOLOX	95.83	95.08	95.24	80.70	95.46
YOLOX+GN	100	89.62	91.77	71.64	94.52
YOLOX+VAFA	100	86.89	88.74	65.71	92.99
YOLOX+GN+VAFA	100	88.52	90.91	69.57	93.91
TJH test set					
YOLOX	77.33	96.0	82.0	98.31	85.66
YOLOX+GN	90.67	80.0	88.0	93.15	85.0
YOLOX+VAFA	91.33	76.0	87.5	91.95	82.96
YOLOX+GN+VAFA	92.0	76.0	88.0	92.0	83.24

Table 3.2: Evaluation results of ablation experiments at the image level using the MCH and TJH test sets.

Model	Sensitivity(%)	Specificity(%)	Accuracy(%)	Precision(%)	F_1-score(%)
MCH test set					
YOLOX	96.01	92.23	92.70	61.39	94.11
YOLOX+GN	98.02	85.54	86.96	46.45	91.36
YOLOX+VAFA	98.68	84.32	85.89	43.69	90.94
YOLOX+GN+VAFA	98.73	84.32	85.95	44.61	90.96
TJH test set					
YOLOX	82.07	91.38	85.66	93.81	86.47
YOLOX+GN	89.25	72.69	82.86	83.88	80.12
YOLOX+VAFA	92.91	70.73	84.40	83.59	80.32
YOLOX+GN+VAFA	92.92	70.41	84.24	83.34	80.11

Table 3.3: Comparison between our proposed model and the EfficientDet model at the image-level and video-level.

Model(video-level)	Sensitivity(%)	Specificity(%)	Accuracy(%)	Precision(%)	F_1-score(%)
MCH test set					
EfficientDet	52.08	97.81	88.31	86.21	67.97
YOLOX+GN+VAFA	100	88.52	90.91	69.57	93.91
TJH test set					
EfficientDet	50.0	98.36	88.31	88.89	66.30
YOLOX+GN+VAFA	92.0	76.0	88.0	92.0	83.24

Model(image-level)	Sensitivity(%)	Specificity(%)	Accuracy(%)	Precision(%)	F_1-score(%)
MCH test set					
EfficientDet	83.16	89.34	88.82	43.51	86.14
YOLOX+GN+VAFA	98.73	84.32	85.95	44.61	90.96
TJH test set					
EfficientDet	79.23	89.93	88.97	44.63	84.24
YOLOX+GN+VAFA	92.92	70.41	84.24	83.34	80.11

3.5 Discussion

In this study, the CAD subsystem based on enhanced YOLOX was applied to the detection of EP in hysteroscopic videos.

It was found that the proposed model could achieve high sensitivity and real-time performance. Compared with the original YOLOX model, the proposed model had improved per-lesion sensitivities using both the internal test set from the MCH and the external test set from the TJH, with 100% and 92.0%, respectively. In particular, the per-lesion sensitivity of 92.0% with the external test dataset was significantly superior to the original model's per-lesion sensitivity of 77.33%, demonstrating the superior generalization capability of the proposed model. This can be attributed to the utilization of the GN method instead of the BN method. BN is strongly dependent on batch size, causing severe variations in the parameter updates and a noticeable decrease in recognition accuracy when small batch sizes are used for training neural networks. Conversely, GN is independent of batch size, leading to consistent performance even with small batch sizes. The integration of the VAFA algorithm further enhances the sensitivity of the proposed model.

In addition, we compared our proposed method with the highly efficient EfficientDet model [21], and the comparison results are shown in Table 3.3. Table 3.3 demonstrates that our proposed model yields superior performance to EfficientDet. Specifically, on the MCH test set, our proposed model achieved video-level sensitivity of 100% and image-level sensitivity of 98.73%, outperforming EfficientDet's corresponding sensitivity of 52.08% and 83.16%, respectively. On the TJH test set, our proposed model also outperformed EfficientDet in terms of sensitivity, achieving video-level sensitivity of 92.0% and image-level sensitivity of 92.92%, compared to EfficientDet's corresponding sensitivity of 50.0% and 79.23%, respectively. Our proposed model also achieved high accuracy and F1-score values compared to EfficientDet. The details are also listed in Table 3.3.

As shown in Tables 3.1 and 3.2, the accuracy of the proposed model was worse than that of the original model on the internal test set from the MCH, but similar or even

better on the external test set from the TJH. We hypothesize that this issue occurred because all models were trained only with data provided by the MCH, whereas the test set comprised data from both hospitals. Considering variations in the data collection equipment and procedures employed by the two hospitals, there could be an inconsistent distribution between the source and target data. In future work, we will adopt domain generalization methods to address this issue.

The current use of deep learning in hysteroscopic images is mainly focused on the classification of endometrial cancer, with a few studies on endometrial fibroids [80,97]. For example, Török et al. used a fully convolutional CNN to identify the plane between myoma and the normal myometrium, achieving a pixel-wise segmentation accuracy of 86.19% after training the network on 13 cases of video data for 140 epochs [**?**]. Deep learning research on uterine lesions has mainly focused on MRIs and ultrasound images. Zhang et al. trained and evaluated the LeNet-5 neural network using MRIs from 158 patients with endometrial cancer, achieving an area-under-the-curve value of 0.897 [101]. Dong et al. applied a U-Net neural network to MRI scans, seeking the depth of endometrial cancer invasion and achieving a model accuracy of 79.2%, which was not significantly different from the diagnostic accuracy achieved by radiologists [102]. Xia et al. used a DPA-UNet neural network to integrate hysteroscopy and ultrasonography for the detection of endometrial cancer [103]. Wang et al. achieved automatic endometrial segmentation and thickness measurements for ultrasound images using a 3D U-Net network [79]. Dilna et al. used the MBF-CDNN method to detect uterine fibroids in ultrasound images [104]. Several studies have investigated MRIs of endometrial fibroids using deep-learning-based methods [105–108]. Overall, these studies demonstrate the potential of deep learning for improving the detection of uterine lesions. The improved YOLOX model developed in this study has shown high sensitivity in the detection of endometrial polyps in hysteroscopic images, and it may be a useful tool for assisting doctors in clinical diagnosis.

To the best of our knowledge, this study is the first to use a deep-learning-based method to detect endometrial polyps based on hysteroscopic images. An improved

YOLOX model was used to achieve real-time detection and high accuracy, making it suitable for use in clinical hysteroscopic surgery. In addition, our VAFA algorithm was proposed for the post-processing stage, with the aim of making the object detection box more stable and reducing discomfort for doctors. Our improved YOLOX model was able to achieve its best performance in fewer iterations and with less time overhead compared with the original YOLOX model. Tables 3.1 and 3.2 show that the proposed model demonstrated significantly higher sensitivity than the original model, particularly with the external test set from the TJH. This is advantageous because the proposed model minimizes the rate of missed detections by doctors. Although the proposed model has lower specificity than the original model, it remains acceptable. Overall, these results demonstrate the potential of deep learning for improving the detection of endometrial polyps in hysteroscopic images.

The limitations of this study can be listed as follows:

1. The model showed poor performance in detecting hysteroscopic images with polyps partially occluded by the endometrium.

2. A large floating endometrium can be misdiagnosed as a polyp. We expect that problems 1 and 2 can be addressed by increasing the number of occluded polyp images and background images in the training set.

3. Deep-learning-based object tracking algorithms, such as Deep SORT, have been employed to address the problem of unsteady detection boxes [109]. However, their performance should be improved further. The proposed VAFA algorithm should also be updated, because the object detection display is insufficiently smooth. Therefore, further research is needed to develop more advanced algorithms that would improve the polyp detection performance.

4. A prospective study should be conducted to check that the proposed method performs as expected in real clinical practice.

In this study, we improved the YOLOX model to achieve higher sensitivity in the detection of endometrial polyps. The VAFA algorithm was also proposed in a post-

processing stage to improve the stability of the detection process and enhance its convenience of use by hysteroscopists. The improved model and method showed significant generalizability and stable capacity and could achieve high sensitivity in the detection of endometrial polyps.

Future works will mainly be based on the practical application of this study. Firstly, we expect to explore the feasibility of integrating our algorithm into existing clinical hysteroscopy systems to improve the efficiency and accuracy of endometrial polyp detection. Secondly, we want to optimize the proposed model by combining hysteroscopic videos and medical information for use on mobile devices or web-based platforms. This would facilitate the usage of the proposed method in remote healthcare services or telemedicine systems.

Overall, the improved model may be useful in reducing the missed diagnosis rate in clinical hysteroscopic surgery and improving the detection sensitivity of endometrial polyps.

3.6 Conclusion

In conclusion, this study proposed a significant application of deep learning methods for detecting endometrial polyps from hysteroscopic images. The introduction of an enhanced YOLOX model facilitated real-time detection with remarkable accuracy. Additionally, the VAFA algorithm was developed for post-processing to optimize object detection. However, certain limitations were identified, including the need for further enhancement of the algorithm for better detection box stability, and further validation of the model in real-world clinical scenarios.

Notwithstanding, the enhanced YOLOX model along with the VAFA algorithm demonstrated exceptional generalizability and stability, rendering them potentially advantageous tools for increasing the detection sensitivity of endometrial polyps. Future research will aim to incorporate these algorithms into current clinical hysteroscopy systems, to augment both efficiency and accuracy of endometrial polyp detection. Fur-

thermore, it is anticipated that the model will be optimized for use with hysteroscopic videos and patient data on mobile and web-based platforms, expanding its applicability to remote healthcare services and telemedicine. Consequently, this study's developments present promising prospects in reducing missed diagnoses in clinical hysteroscopic surgery and improving detection sensitivity for endometrial polyps.

Chapter 4

Computer-Aided Diagnosis of Endometrial Cancer and Atypical Endometrial Hyperplasia Based on Deep Learning

4.1 Introduction

EC is a common gynecological cancer that occurs in the endometrium [11]. EC is rapidly increasing in women in high-income countries [12].EC has become the fourth prevalent female cancer in high-income countries [13]. In 2020, the world witnessed 417,367 new cases and 97,370 fatalities attributed to the condition [14]. Although EC is generally considered to have a favorable outcome, high-grade cancers tend to recur and a more advanced stage or in metastatic lesions, posing a significant therapeutic challenge [15]. AEH is a precancerous lesion of the endometrium. After hysterectomy, EC was reportedly diagnosed in 27% to 52% of patients who had preoperative AEH [16]. The risk of concurrent endometrial cancer in patients with AEH during hysterectomy is

40% [17]. Accurate diagnosis of EC and AEH is crucial for early treatment, given the rapid progression of these lesions.

Experienced surgeons often fail to curette more than half of the uterine cavity in 60% of D&C procedures, which can complicate the diagnosis, particularly in instances of focal uterine lesions [110,111]. For the direct evaluation and detection of structural abnormalities in the endometrium and uterine cavity, hysteroscopy has been generally accepted as a safe and minimally invasive method [112]. With the rapid development in endoscopic technology, endometrial cancer surgery can be diagnosed through hysteroscopy, which is an effective tool.

However, hysteroscopy, relying solely on the subjective judgment of the hysteroscopists, can introduce inherent variability and potential errors in the diagnostic process [18]. Inexperienced hysteroscopists may face challenges in accurately identifying and interpreting abnormalities, leading to decreased diagnostic accuracies. This, in turn, can result in delays in initiating timely and appropriate treatment for patients, causing suboptimal outcomes and potential economic repercussions. Therefore, there is a critical demand for interventions focused on augmenting the objectivity and precision of hysteroscopy, notably through the incorporation of CAD systems. These interventions aim to alleviate the challenges posed by limited clinical expertise, thereby reducing the risks associated with misdiagnosis, treatment delays, and financial implications.

4.2 Outline

In this section 4.1, we introduce EC as a prevalent gynecological malignancy affecting the endometrium. We highlight the increasing incidence of EC in high-income countries and the challenges posed by high-grade cancers and advanced or metastatic lesions. The importance of accurate diagnosis for early treatment, particularly in the case of AEH, is emphasized. We also discuss the subjective nature of hysteroscopy and the need for computer-aided diagnostic systems to enhance objectivity and precision, reducing misdiagnosis risks, treatment delays, and associated economic implications.

In section 4.3, we review the application of deep learning techniques in endometrial cancer image analysis. We introduce the success of neural networks such as LeNet-5, U-Net, and VGGNet-16 in tasks such as endometrial cancer classification, depth estimation, and lesion localization in hysteroscopic images. we discussed the utilization of attention mechanisms to enhance the attention capabilities of CNN in computer vision tasks. These mechanisms enable models to selectively focus on important regions and features and improve comprehension of visual content.

Section 4.4 provides a comprehensive description of our research method, beginning with the description of the data collected. This is followed by a detailed discussion on the preprocessing of the hysteroscopy images for the model. The preprocessing steps include isolating regions of interest and standardizing the dimensions of the images through strategic cropping and resizing to ensure the model's focus on crucial areas. The section then elucidates the effectiveness of the EfficientNet neural network, further emphasizing the augmentation achieved through the integration of ParNet attention. To address the class representation disparity in the dataset, a class weighting strategy is adopted. This strategic implementation enhances model performance and robustness, consequently promoting a balanced consideration of all classes.

This section 4.5 reports on the development of the model, detailing the network configuration, such as the optimizer used, learning rate, momentum, and decay rate. The model's effectiveness is also evaluated through the AUC value, which demonstrated the improvements achieved through the implementation of ParNet attention and class weighting.

The final section 4.6 elaborates on the study's overall findings, highlighting the model's success in improving the classification of EC/AEH lesions. The potential for the model to aid in the diagnosis of endometrial cancer is also proposed, with future considerations towards evaluating the study's efficacy in practical clinical applications.

4.3 Literature Review

Deep learning has been widely implemented in endometrial cancer image analysis. Zhang et al. [113] utilized MRIs from 158 patients with endometrial cancer to train and evaluate the performance of the LeNet-5 neural network. The model achieved an impressive AUC value of 0.897. Similarly, Dong et al. [101] developed and evaluated a U-Net neural network approach to determine the depth of endometrial cancer invasion using MRI scans. The model achieved a high accuracy of 79.2%. Urushibara et al. [114] employed the Xception model for the diagnosis of endometrial cancer through the analysis of MRI data. Some researchers [115] employed VGGNet-16 for endometrial lesion classification in hysteroscopic images. Another study [97] proposed a deep learning-based system for the automatic localization of endometrial cancer.

Attention mechanisms are essential computational techniques in computer vision models, allowing them to focus on specific parts of the input data based on varying levels of importance assigned to different elements. These mechanisms play a crucial role in improving performance and enhancing understanding of visual content. In recent years, Woo et al. [116] proposed the Convolutional Block Attention Module (CBAM) to enhance the attention capabilities of Convolutional Neural Networks in image classification. In contrast, another approach known as the Bottleneck Attention Module (BAM) [117] introduces a different design. The BAM module incorporates a bottleneck structure that reduces computational complexity and memory requirements, making it more efficient while preserving attention capabilities. This module aims to strike a balance between performance and resource utilization, offering advantages over CBAM in terms of efficiency. Additionally, the Squeeze-and-Excitation (SE) module, proposed in [118], introduces a distinct attention mechanism. This SE module effectively enhances CNNs by selectively amplifying important channels and suppressing less relevant ones.

To summarize, attention mechanisms, such as the CBAM, BAM, and SE modules, have been proposed to enhance the attention capabilities of CNNs in computer vision tasks. These modules enable models to selectively focus on important regions and fea-

tures in the input data, leading to improved performance, a better understanding of visual content, and efficient resource utilization.

4.4 Methods

4.4.1 Dataset

The hysteroscopy images were collected from the Maternal and Child Hospital of Hubei Province (MCH) in 2008-2019. Table 4.1 displays the data distribution for the different categories in both the training and test sets. The categories are divided into ”EC/AEH” and ”Control” groups. In the training set, there are 131 cases belonging to the EC/AEH category, with a total of 3,122 images. For the control group, there are 1,106 cases with 46,434 images. The test set includes 23 cases from the EC/AEH category, comprising a total of 698 images. The control group in the test set consists of 62 cases and 2,714 images. The control group consisted of normal uterine cavities, uterine leiomyoma, endometrial polyps, and endometrial hyperplasia without atypia. Table 4.2 lists the detailed information of the control group. Pathologists diagnosed all lesions depicted in the pathological images. This study was approved by the Medical Ethics Committee of MCH. The training and analysis were conducted anonymously to comply with the privacy policy.

Table 4.1: The datasets for EC/AEH classification

Category	Training set		Test set	
	Cases	Images	Cases	Images
EC/AEH	131	3,122	23	698
Control	1,106	46,434	62	2,714

Table 4.2: Detail of Control group.

Category	Training set		Test set	
	Cases	Images	Cases	Images
NE	499	10,977	41	2,151
P	260	9,166	21	563
UL	194	19,866	-	-
EH	153	6,425	-	-

4.4.2 Data Preprocessing

In the process of preparing our hysteroscopy images for the model, we implemented several steps to isolate the regions of interest and enhance the robustness of our dataset. This involved cropping the images to eliminate any non-lesion areas that were not relevant to our study, thereby focusing the model's attention on the critical areas of each image. The images were then subjected to a series of transformations to standardize their dimensions and introduce variability. Initially, the images were resized to a uniform dimension of 224×224 pixels. This was followed by center cropping, a technique that ensures consistency in image dimensions across the dataset, which is crucial for the model to process the images effectively.

To increase the robustness of our model and its ability to generalize to unseen data, we introduced variability into our dataset through random affine transformations. These transformations included a translation of up to 0.05 in each direction and a random horizontal flip. These techniques serve to augment the dataset by creating variations of the original images, thereby expanding the range of data the model is exposed to during training. Finally, the augmented images were converted into tensor format, a data structure that is compatible with the PyTorch framework and efficient for computational operations. The images were then normalized using mean and standard deviation values of [0.485, 0.456, 0.406] and [0.229, 0.224, 0.225], respectively. Normalization is a critical step in preparing data for neural networks, as it ensures that the range of pixel values is consistent across images and helps to improve the stability and performance

of the model.

4.4.3 EfficientNet Neural Network

In previous convolutional neural networks, the accuracy of the network was often improved by increasing the image resolution, network width, or network depth individually. However, as the model size increased, its accuracy would often decrease. To address this challenge, Tan et al. [2] introduced a standardized approach for scaling and balancing different dimensions of the network. They used the Neural Architecture Search (NAS) technique to design the EfficientNet network, which introduced a novel model scaling method using a simple and efficient compound coefficient. The new method provides a new paradigm for model scaling that achieves enhanced performance while maintaining efficiency.

In this study, we propose an EC/AEH computer-aided diagnosis method based on EfficientNet-B0. EfficientNet-B0 is a baseline model in the EfficientNet architecture. The network structure is shown in Figure 4.1. Stage 1 is a convolutional layer with a convolution kernel size of 3×3. Stage 2-8 are repeatedly stacked the Mobile Inverted Bottleneck Convolution (MBConv) modules, and Stage 9 consists of a 1×1 convolutional layer, an average pooling layer, and a fully connected layer.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	28×28	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Figure 4.1: EfficientNet-B0 network structure [2].

4.4.4 ParNet Attention

The ParNet Attention [119] is an attention mechanism, integrating elements of spatial and channel-wise feature learning with a modified SE block, termed as SSE. The fundamental concept in this structure is the ability to capture both channel-wise and spatial information simultaneously, which is enabled by two separate convolutional layers. The ParNet structure is shown in Figure 4.2. A 1×1 convolutional layer is utilized for channel-wise feature extraction, while a 3×3 convolutional layer is employed for spatial context acquisition. The SSE block is a critical aspect of the ParNetAttention module, designed to overcome the limited receptive field challenge of non-deep networks with only 3×3 convolutions. The SSE design is based on the traditional SE network but features certain key modifications to better suit our objectives.

Unlike the vanilla SE design, which increases the depth of the network, the SSE mechanism is applied alongside the skip connection, ensuring the network depth remains manageable. The SSE block uses Adaptive Average Pooling and a 1×1 convolution to capture global spatial information and channel-wise dependencies, respectively. The outputs from the convolutional layers and the SSE block are aggregated and passed through a Sigmoid Linear Unit (SiLU) activation function. This function introduces the desired level of non-linearity into the model, enabling it to model complex patterns effectively. The ParNet Attention module thus provides a comprehensive and nuanced attention mechanism, simultaneously capturing spatial and channel-wise features while also embedding global information through the SSE block. Its design ensures a balance between complexity and computational efficiency.

The integration of the ParNet Attention module within the EfficientNet-B0 architecture offers notable benefits over the traditional SE attention module. The ParNet Attention module, incorporating the SSE block, achieves this without increasing the network's complexity, resulting in a more resource-efficient model. By replacing the SE module with the ParNet Attention module, our model acquires a more nuanced feature representation, capturing spatial and channel-wise information simultaneously. This enhancement improves the model's comprehension of hysteroscopy images, leading to

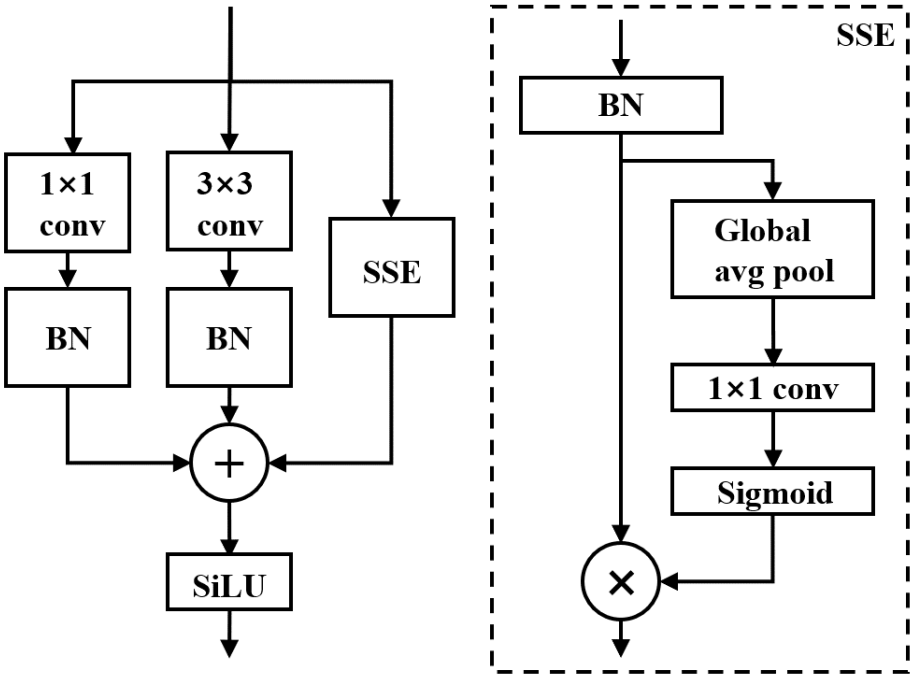


Figure 4.2: ParNet attention mechanism.

comparable performance while preserving the network depth. The MBConv module in our model is illustrated in Figure 4.3.

4.4.5 Class Weighting

Our training dataset comprises 3,122 images classified as EC/AEH and a significantly larger set of 46,434 images classified as Control. This disparity in class representation presents a challenge, as it can lead to a model that is biased toward the majority class. To mitigate this issue during the training of our EfficientNet model, we implemented a class weighting strategy.

The class weighting strategy involves assigning weights to each class in inverse proportion to their representation in the dataset. This means that the EC/AEH class, which is underrepresented, is assigned a higher weight compared to the Control class. These weights are then incorporated into the cross-entropy loss function, which is used to train the model. The effect of this weighting is to increase the penalty for misclassifying

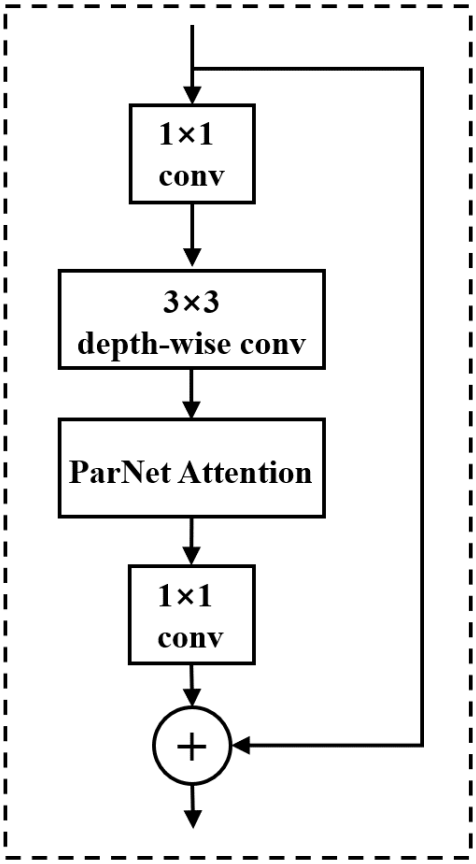


Figure 4.3: MBConv module in EfficientNet-B0 based on ParNet attention mechanism.

Table 4.3: Evaluation results of ablation experiments using the test set.

	EfficientNet-B0	+ParNet attention	+Class weighting	+ParNet attention +Class weighting (Our model)
AUC(95% CI)	0.886(0.797-0.974)	0.923(0.858-0.988)	0.930(0.860-1.0)	0.941(0.891-0.990)
Accuracy (95% CI)	89.4%(89.2–89.6%)	88.2%(88.0-88.5%)	83.5%(83.2-83.8%)	89.4%(89.2-89.6%)
Sensitivity	82.6%(67.1-98.1%)	89.9%(79.8-1.0%)	89.9%(79.8-1.0%)	93.7%(87.3-1.0%)
Specificity	91.9% (85.2-98.7%)	87.1%(78.8-95.4%)	90.3%(83.0-97.7%)	87.1%(78.8-95.4%)
PPV (95% CI)	79.2%(62.9-95.4%)	72.4%(56.1-88.7%)	77.8%(62.1-93.5%)	73.3%(57.5-89.2%)
Kappa	0.735(0.573-0.898)	0.725(0.567-0.882)	0.774(0.626-0.922)	0.755(0.607-0.903)
F_1-score	0.8086	0.8313	0.8341	0.8225

EC/AEH images, thereby making the loss function more sensitive to prediction errors in the EC/AEH class. This encourages the model to pay more attention to the EC/AEH class during training, despite its smaller sample size. By implementing this approach, we aimed to address the class imbalance in our dataset and improve the model's performance in detecting and classifying EC/AEH images. This is a critical step towards ensuring that our model is robust and performs well across all classes, regardless of their representation in the training data. The specific method for calculating the weights is as follows:

- For each class, calculate the reciprocal of the sample count to obtain a weight factor. For class EC/AEH, the weight factor is 1/3211; for class control, the weight factor is 1/46434.

- Normalize the weight factors so that their sum equals 1. Specifically, divide each weight factor by the sum of all weight factors to obtain normalized weights. For class EC/AEH, the normalized weight is 0.935; for class control, the normalized weight is 0.065.

- Utilize the normalized weights as class weights in the calculation of the cross-entropy loss function.

4.4.6 Evaluation Metrics

We utilize the classification performance of the model at the case level as a benchmark for comparing and evaluating the performance of different models. The predicted malignant score for each case was calculated following the approach of Zhou et al. [120]. The formula is shown in equation 1:

$$\begin{aligned}\theta = &- [w_1 \times \log_{10}(1 - p_1) + [w_2 \times \log_{10}(1 - p_2) \\ &+ \cdots + [w_n \times \log_{10}(1 - p_n)]/n\end{aligned} \tag{4.1}$$

where n is the total number of images from the case, $P_{malignancy} = [p_1, p_2, \ldots, p_n]$ is the predicted probabilities of these n images classified as malignancy. w_1 is calculated as $w_i = p_i/(p_1 + p_2 + \cdots + p_n)$.

The classification performance of the models was evaluated using ROC curves, AUC, accuracy, sensitivity, specificity, PPV, F_1-Score Kappa, and related 95% confidence interval (CI). The evaluation metrics employed in this chapter align with the computational method described in Section 2.4.3. All values for the evaluation metrics were computed using the reportROC package (version 3.5) in the R programming language.

4.5 Results

The proposed model was developed with Pytorch framework. A Stochastic Gradient Descent optimizer was used with a learning rate of 0.01, a momentum of 0.9, and a decay rate of 4e-4. The training epoch is set to 50 and batch size is set to 20. The model was configured with a scaling factor of 1.0 for both the channel dimension and the depth dimension. Within the MBConv module, the dropout layer was set with a dropout rate of 0.2 to introduce random dropout. The dropout rate of 0.2 was applied before the last fully connected layer of the model.

To assess the efficacy of the improvements to the Efficient-B0 model, we applied it in ablation experiments. Table 4.3 shows a detailed comparison of the classification performance achieved by different models on the test set. Our proposed model, incorporating ParNet attention and class weighting modifications, achieved the highest AUC of 0.941 compared to other models with AUCs of 0.886, 0.923, and 0.930. This indicates the superior discriminative ability of our model in accurately classifying EC/AEH from benign lesions. Specifically, our proposed model reached accuracy of 89.4%, sensitivity of 93.7%, specificity of 87.1%, PPV of 73.3%, Kappa of 0.755, and F_1-Score of 0.82. These results demonstrate the efficiency of combining ParNet attention and class weighting techniques to enhance EfficientNet-B0 performance. The ROC curves

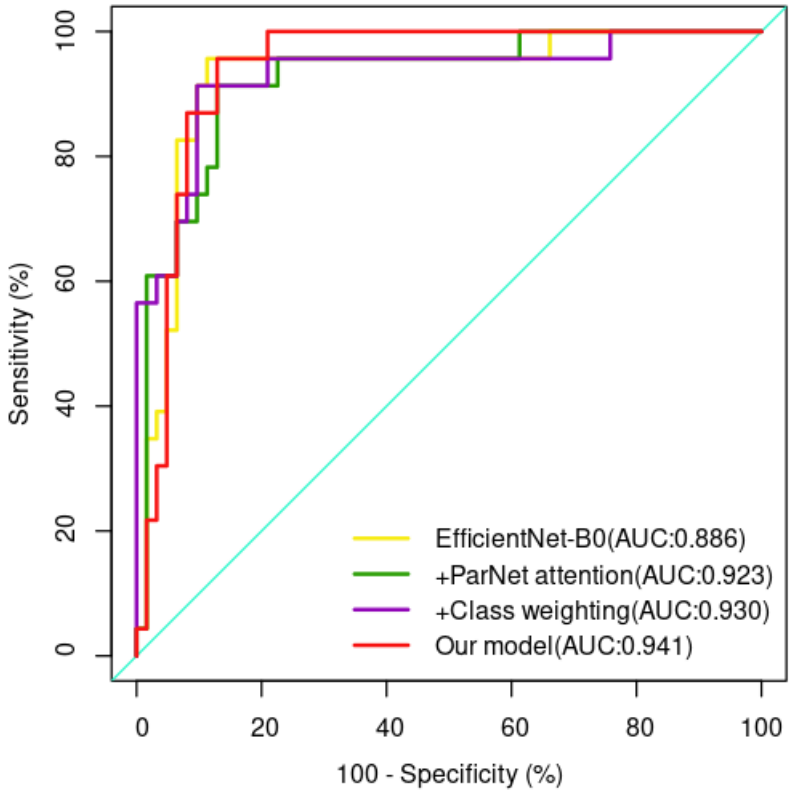


Figure 4.4: ROC curves of the models.

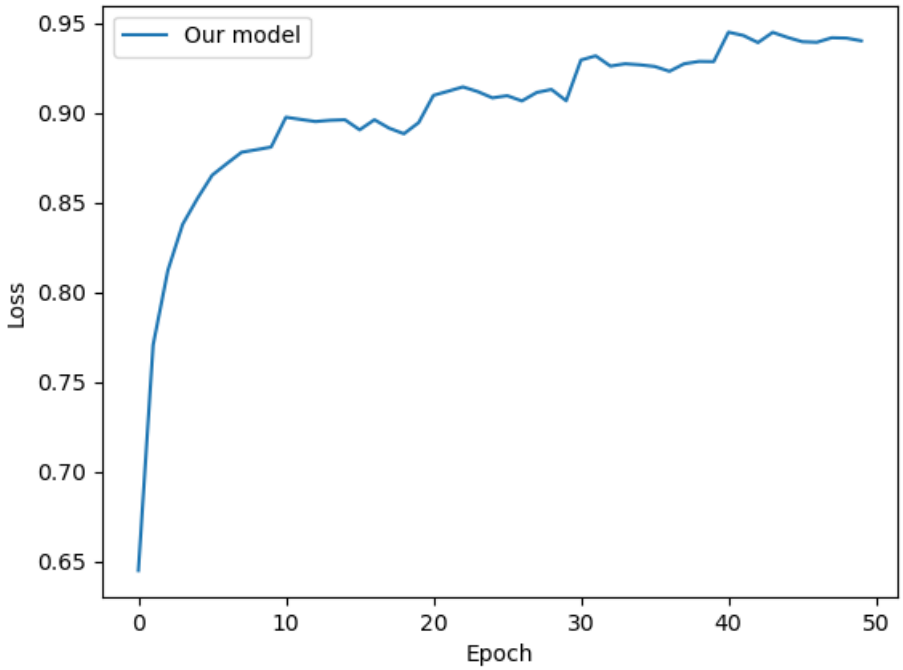


Figure 4.5: The Loss value curve of our model during training.

of different models are shown in Figure 4.4. Figure 4.5 shows the loss curve of our proposed model. Some examples of the classification results of our proposed model on the test set are displayed in Figure 4.6.

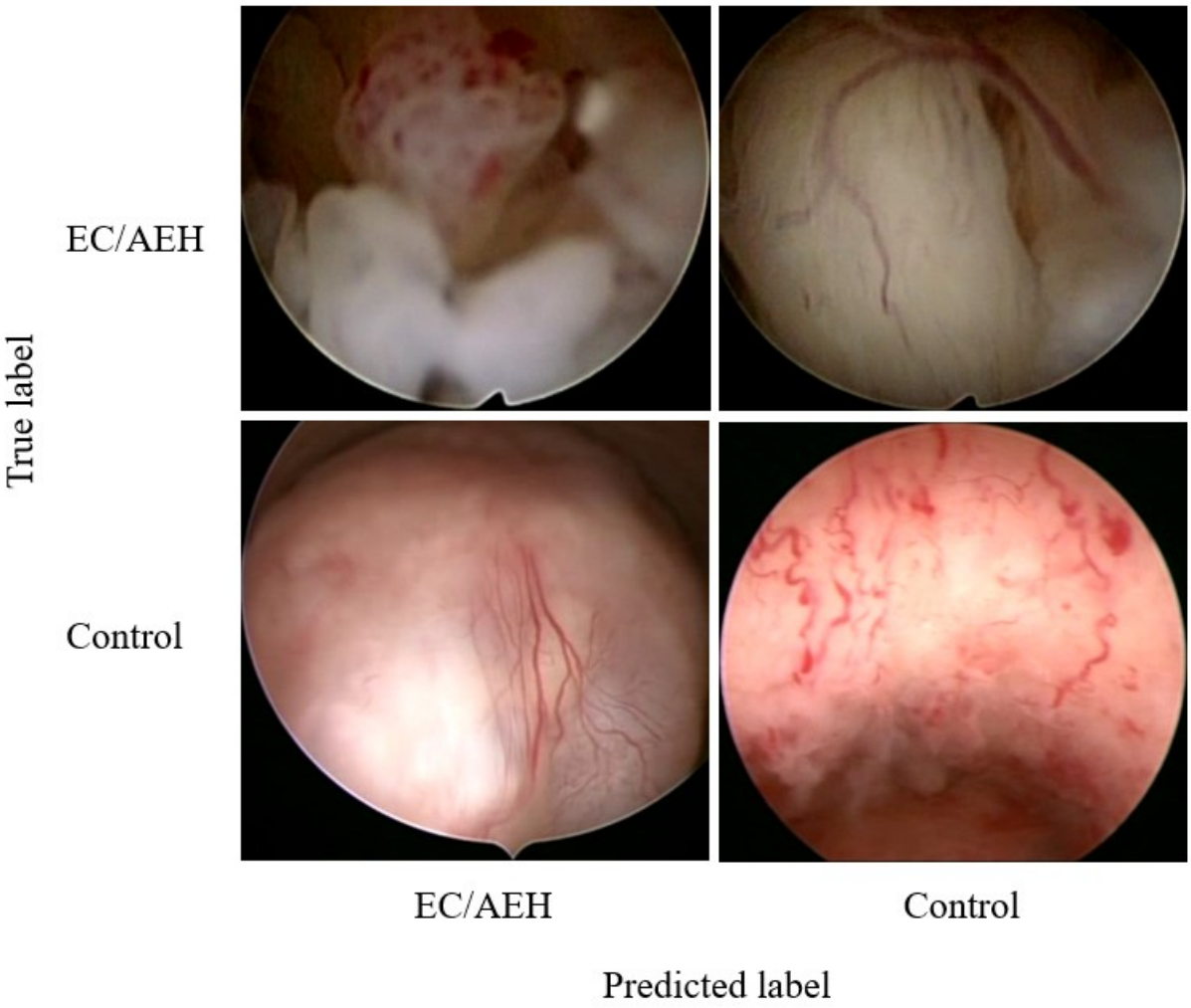


Figure 4.6: Example classification results output by our model.

4.6 Discussion

In this comprehensive investigation, we embarked on an endeavor to augment the capabilities of the EfficientNet-B0 model by incorporating a sophisticated attention mechanism known as ParNet and leveraging the concept of class weighting to balance the influence of different classes. The EfficientNet-B0 model, a versatile and potent architecture in the field of machine learning, serves as an optimal starting point due

to its compactness and efficiency. By integrating the ParNet mechanism, we aimed to refine the model's feature selection process, consequently enhancing its predictive accuracy. The adoption of class weighting further promoted equitable class representation, countering potential class imbalance issues that could impair the model's generalization performance.

The cardinal objective of our research revolved around designing and optimizing a CAD subsystem specifically tailored for the categorization of EC and AEH lesions. These conditions have substantial implications for women's health on a global scale, warranting robust detection and classification systems. The CAD subsystem, often a key component in modern healthcare, has the potential to offer timely and accurate diagnostic support to clinicians, thereby facilitating earlier intervention and improved patient outcomes.

Our primary aim was to devise a model that exhibits exceptional discrimination capacity, being able to distinguish between EC and AEH lesions effectively and reliably. Our proposed model achieved an AUC of 0.941. This robust score underscores the model's superior capacity for binary classification, demonstrating high sensitivity and specificity. Sensitivity and specificity are pivotal metrics in medical diagnosis.

The primary innovation of our approach is the integration of ParNet attention within the EfficientNet-B0 model. By doing so, we aimed to enhance the model's capacity to accurately classify lesions, recognizing the pivotal role that attention mechanisms can play in focusing on critical image features. ParNet attention, in particular, has demonstrated effectiveness in tasks requiring the interpretation of complex image data, making it a fitting choice for integration into our model.

Alongside this, we adopted class weighting as part of our strategy to address potential class imbalance issues. Such imbalances are commonplace in medical image datasets and can skew the training process of machine learning models, leading to the overrepresentation of one class at the expense of the other. The application of class weighting assists in normalizing this disparity, ensuring that both classes are duly represented during the model's learning process.

Our proposed model's utility extends beyond the research context, potentially having significant implications in the clinical diagnosis of endometrial cancer and other non-cancerous conditions. Its high AUC indicates its potential in differentiating between EC/AEH lesions, suggesting its usefulness as a diagnostic tool.

Future work will be centered around an in-depth evaluation of our model's effectiveness within practical clinical scenarios, particularly within the context of the CAD subsystem. This step is deemed essential given the model's notable performance in our initial study. However, it's also critical to comprehend the applicability of the model in real-world scenarios, where factors like data variability and interpretability could substantially influence its performance. In order to validate its utility robustly, we will employ diverse validation strategies. These will range from conducting retrospective analyses on historical patient data to implementing prospective trials. This approach will facilitate a comprehensive assessment of the model's clinical relevance and its potential to contribute significantly to CAD systems.

4.7 Conclusion

In this comprehensive study, we enhanced the capabilities of the EfficientNet-B0 model through the integration of the ParNet attention mechanism and the adoption of class weighting strategies, aiming to optimize a CAD subsystem for the classification of endometrial cancer and atypical endometrial hyperplasia. Our model, with AUC of 0.941, demonstrated superior classification capacities with high sensitivity and specificity, hence presenting a promising diagnostic tool.

The primary innovation was the embedding of ParNet attention within the EfficientNet-B0 architecture, intended to enhance the accurate classification of lesions by emphasizing critical image features. Additionally, we addressed potential class imbalance issues. This strategy ensures equitable representation of all classes during the model's training process, thereby preventing any overrepresentation.

Our model holds potential clinical implications in the diagnosis of endometrial can-

cer and other related conditions. Future work will focus on an extensive assessment of the CAD subsystem's effectiveness in practical clinical scenarios.

Chapter 5

Conclusion

Our study presents a novel CAD system for the recognition of UF, the detection of EP, and the diagnosis of EC and AEH from various lesions.

In the realm of UF recognition, we have successfully developed a CAD subsystem that combines CNN and Transformer models. Our proposed hybrid model demonstrates superior performance with an accuracy of 0.8893, surpassing existing models and showcasing its effectiveness in accurately identifying UF.

For EP detection, our enhanced CAD subsystem, incorporating the YOLOX model and the VAFA algorithm, exhibits notable improvements in sensitivity. The augmented model demonstrates robust generalizability and stability, leading to heightened sensitivity in EP detection. This advancement has the potential to significantly reduce missed diagnosis rates during clinical hysteroscopic surgery.

Additionally, we introduce another CAD subsystem that integrates ParNet module and class weighting method to enhance the performance of the EfficientNet-B0 model in classifying EC/AEH lesions. The proposed model achieves a higher AUC value of 0.941, underscoring its potential for accurate diagnosis of both EC and non-cancerous lesions.

The comparative analysis of our CAD system and other relevant system is visually represented in Figure 5.1. In fact, there is a dearth of research involving the application of CAD systems to uterine lesion images. The systems depicted in the figure did not

utilize identical datasets, yet the illustration still offers insights. It is evident that our CAD system closely aligns with the current mainstream CAD systems currently utilized in CAD diagnostic systems based on uterine lesion imagery.

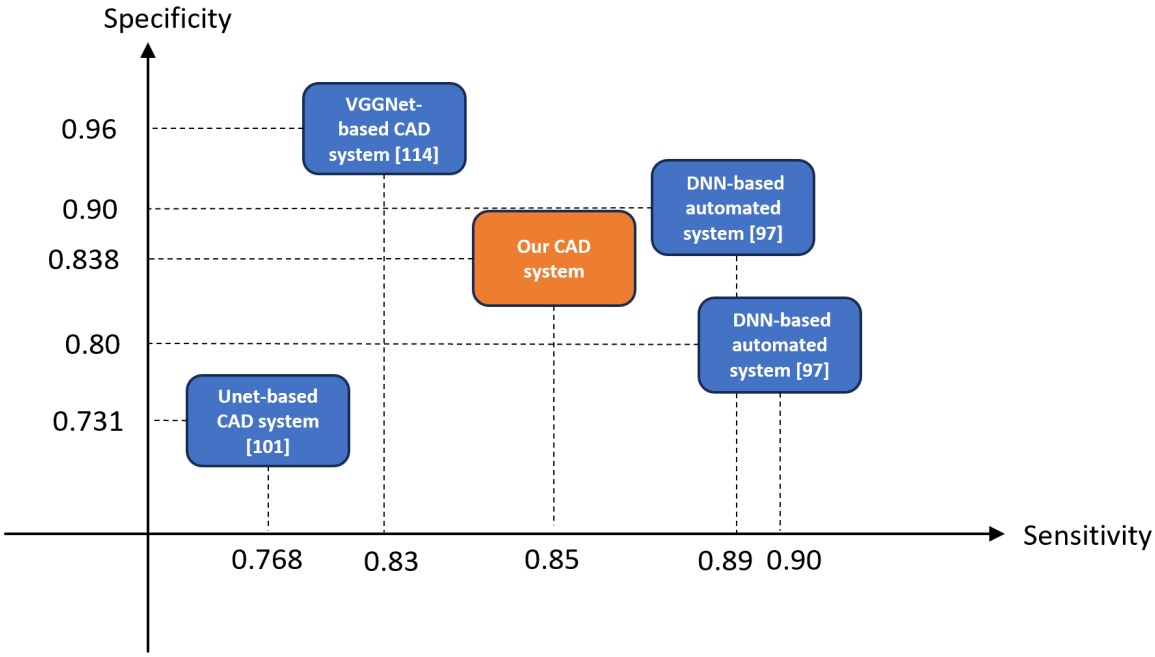


Figure 5.1: Comparison of our CAD system with other related systems.

5.1 Limitations

This dissertation faces several limitations that are necessary to note.

In the current stage of our research, we have yet to fully integrate the various CAD subsystems that respectively perform the tasks of recognizing UF, detecting EP, and classifying EC/AEH. Each subsystem contributes significant value to the overall CAD system, serving an important function in the analysis and diagnosis process.

However, while each of these subsystems provides valuable assistance in its respective area, their current standalone status makes them less efficient than they could potentially be. The lack of integration between these subsystems means that healthcare professionals must use each one independently, which can lead to a complex and cumbersome process. This fragmentation may not only increase the time taken to complete

the diagnostic process but could also lead to misinterpretations or oversights, particularly in scenarios where the conditions overlap or present together.

Therefore, the integration of these subsystems into a unified CAD system is a crucial next step in our research. Not only would this make the system more efficient and user-friendly, but it would also provide a more holistic view of a patient's health status, enabling healthcare providers to diagnose and treat multiple conditions in a more streamlined and effective manner. Integration of these systems represents a significant opportunity for improvement, enhancing the overall workflow and potentially improving patient outcomes.

In addition, the Specific limitation is the CAD subsystem for the detection EP showed compromised performance when detecting polyps partially occluded by the endometrium. A scenario like this represents a challenging test case, as the model struggles to differentiate between polyps and the surrounding endometrial tissue when the polyps are not entirely visible.

Similarly, another related issue arose when distinguishing between large floating endometria and polyps. The CAD subsystem had a propensity to misidentify the former as the latter, indicating the need for improved discernment in situations where these features coexist. While these are certain limitations, they are not insurmountable. The model's sensitivity to these scenarios could potentially be improved by enriching the training dataset with more images containing occluded polyps and an extensive array of background images. This would expose the model to a broader spectrum of scenarios, thereby better equipping it to distinguish between different morphological characteristics under various conditions.

5.2 Future Research

In the future, we will integrate these separate subsystems into a comprehensive and unified CAD hysteroscopy system platform. By merging these subsystems, we aim to simplify the use of CAD systems by healthcare professionals and create a versatile and

user-friendly CAD platform to assist healthcare professionals in the accurate diagnosis and interpretation of hysteroscopic images. By merging these subsystems, we seek to simplify the workflow of healthcare professionals and enable them to make well-informed decisions regarding patients.

This overall system will provide efficient and reliable support to improve the diagnostic capabilities of healthcare providers. In addition, there is great potential for integrating the CAD hysteroscopy system into a telehealth service or telemedicine system. It would allow remote access and analysis of hysteroscopic images, enable timely consultations and expert opinions, and even provide remote assistance in areas with less developed medical conditions. Such integration could eliminate geographic barriers and improve healthcare delivery. The amalgamation of these CAD subsystems into a comprehensive platform holds immense potential in improving the diagnostic capabilities of healthcare providers and optimizing patient outcomes. Our research contributes to the advancement of CAD systems in the field of hysteroscopy and offers promising opportunities for the integration of technology into clinical practice

In conclusion, our study introduces a novel CAD system that leverages CNN, transformer, and attention mechanisms for UF recognition, EP detection, and the diagnosis of EC and AEH lesions. The exceptional performance achieved by our CAD subsystems validates their effectiveness in accurate diagnosis, highlighting their potential to enhance clinical decision-making and improve patient care. As we continue to refine and integrate these subsystems, we anticipate significant advancements in the field of CAD hysteroscopy systems.

References

[1] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 367–376.

[2] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[3] G. G. I. M. C. V. D. A. D. S. V. D. F. C. F. G. M. M. P. Marianna Gulisano, Ferdinando Antonio Gulino, "Role of hysteroscopy on infertility: The eternal dilemma," *CEOG*, vol. 50, no. 5, pp. 99–null, 2023.

[4] M. S. Kamath, J. F. Rikken, and J. Bosteels, "Does laparoscopy and hysteroscopy have a place in the diagnosis of unexplained infertility?" in *Seminars in Reproductive Medicine*, vol. 38, no. 01. Thieme Medical Publishers, Inc., 2020, pp. 029–035.

[5] A. Navarro, M. V. Bariani, Q. Yang, and A. Al-Hendy, "Understanding the impact of uterine fibroids on human endometrium function," *Frontiers in cell and developmental biology*, vol. 9, p. 633180, 2021.

[6] L. I. Zepiridis, G. F. Grimbizis, and B. C. Tarlatzis, "Infertility and uterine fibroids," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 34, pp. 66–73, 2016.

[7] Z. Yanhua, "Clinical significance of gynecological screening in early cervical cancer screening," *World's Latest Med. Inf. Dig*, vol. 15, p. 130, 2015.

[8] J. Hong, "Study on the incidence of endometrial polyps in common gynecological diseases," *Xinjiang Medical University: Xin- jiang, China*, 2013.

[9] G. Wenqian, "Clinical observation of hysteroscopy in the treatment of endometrial polyps," *Chin. J. Metall. Ind. Med.*, vol. 38, p. 202–203, 2021.

[10] Y. T. W. Z. H. Xiang, W.; Qi, "Clinicopathological analysis of postmenopausal endometrial polyps," *Chin. J. Obstet. Gyn.*, vol. 56, p. 131–136, 2021.

[11] E. A. Goebel, A. Vidal, X. Matias-Guiu, and C. Blake Gilks, "The evolution of endometrial carcinoma classification through application of immunohistochemistry and molecular diagnostics: past, present and future," *Virchows Archiv*, vol. 472, pp. 885–896, 2018.

[12] J.-Z. Guo, Q.-J. Wu, F.-H. Liu, C. Gao, T.-T. Gong, and G. Li, "Review of mendelian randomization studies on endometrial cancer," *Frontiers in Endocrinology*, vol. 13, 2022.

[13] K. Passarello, S. Kurian, and V. Villanueva, "Endometrial cancer: an overview of pathophysiology, management, and care," in *Seminars in oncology nursing*, vol. 35, no. 2. Elsevier, 2019, pp. 157–165.

[14] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[15] P.-H. Wang, S.-T. Yang, C.-H. Liu, W.-H. Chang, F.-K. Lee, and W.-L. Lee, "Endometrial cancer: part i. basic concept," *Taiwanese Journal of Obstetrics and Gynecology*, vol. 61, no. 6, pp. 951–959, 2022.

[16] M. Kamii, Y. Nagayoshi, K. Ueda, M. Saito, H. Takano, and A. Okamoto, "Laparoscopic surgery for atypical endometrial hyperplasia with awareness regarding the possibility of endometrial cancer," *Gynecology and Minimally Invasive Therapy*, vol. 12, no. 1, p. 32, 2023.

[17] M. H. Vetter, B. Smith, J. Benedict, E. M. Hade, K. Bixel, L. J. Copeland, D. E. Cohn, J. M. Fowler, D. O'Malley, R. Salani *et al.*, "Preoperative predictors of endometrial cancer at time of hysterectomy for endometrial intraepithelial neoplasia or complex atypical hyperplasia," *American journal of obstetrics and gynecology*, vol. 222, no. 1, pp. 60–e1, 2020.

[18] S. van Wessel, T. Hamerlynck, B. Schoot, and S. Weyers, "Hysteroscopy in the netherlands and flanders: a survey amongst practicing gynaecologists," *European journal of obstetrics & gynecology and reproductive biology*, vol. 223, pp. 85–92, 2018.

[19] L. C. L. Y. S. Huihong, H.; Wantao, "Nursing progress of complications of gynecological hysteroscopic surgery," *Chin. Med. Sci.*, vol. 10, p. 65–68, 2020.

[20] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[22] 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.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

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

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] 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.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *Advances in neural information processing systems*, vol. 32, 2019.

[31] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 213–229.

[32] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning*. PMLR, 2020, pp. 1691–1703.

[33] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[34] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[35] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[36] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[37] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, "Benign and malignant breast tumors classification based on region growing and cnn segmentation," *Expert Systems with Applications*, vol. 42, no. 3, pp. 990–1002, 2015.

[38] M. M. Jadoon, Q. Zhang, I. U. Haq, S. Butt, and A. Jadoon, "Three-class mammogram classification based on descriptive cnn features," *BioMed research international*, vol. 2017, 2017.

[39] H. Chougrad, H. Zouaki, and O. Alheyane, "Deep convolutional neural networks for breast cancer screening," *Computer methods and programs in biomedicine*, vol. 157, pp. 19–30, 2018.

[40] K. Kaur and S. Mittal, "Classification of mammography image with cnn-rnn based semantic features and extra tree classifier approach using lstm," *Materials Today: Proceedings*, 2020.

[41] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated cnn approach," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4701–4709, 2021.

[42] L. Sun, H. Sun, J. Wang, S. Wu, Y. Zhao, and Y. Xu, "Breast mass detection in mammography based on image template matching and cnn," *Sensors*, vol. 21, no. 8, p. 2855, 2021.

[43] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[44] J. SM, C. Aravindan, and R. Appavu, "Classification of skin cancer from dermoscopic images using deep neural network architectures," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15 763–15 778, 2023.

[45] D. Keerthana, V. Venugopal, M. K. Nath, and M. Mishra, "Hybrid convolutional neural networks with svm classifier for classification of skin cancer," *Biomedical Engineering Advances*, vol. 5, p. 100069, 2023.

[46] N. ul Huda, R. Amin, S. I. Gillani, M. Hussain, A. Ahmed, and H. Aldabbas, "Skin cancer malignancy classification and segmentation using machine learning algorithms," *JOM*, pp. 1–15, 2023.

[47] S. B. Mukadam and H. Y. Patil, "Skin cancer classification framework using enhanced super resolution generative adversarial network and custom convolutional neural network," *Applied Sciences*, vol. 13, no. 2, p. 1210, 2023.

[48] K. Mridha, M. M. Uddin, J. Shin, S. Khadka, and M. Mridha, "An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system," *IEEE Access*, 2023.

[49] A. H. Thieme, Y. Zheng, G. Machiraju, C. Sadee, M. Mittermaier, M. Gertler, J. L. Salinas, K. Srinivasan, P. Gyawali, F. Carrillo-Perez *et al.*, "A deep-learning algorithm to classify skin lesions from mpox virus infection," *Nature Medicine*, vol. 29, no. 3, pp. 738–747, 2023.

[50] F. Olayah, E. M. Senan, I. A. Ahmed, and B. Awaji, "Ai techniques of dermoscopy image analysis for the early detection of skin lesions based on combined cnn features," *Diagnostics*, vol. 13, no. 7, p. 1314, 2023.

[51] M. Altun, H. Gürüler, O. Özkaraca, F. Khan, J. Khan, and Y. Lee, "Monkeypox detection using cnn with transfer learning," *Sensors*, vol. 23, no. 4, p. 1783, 2023.

[52] H. K. Gajera, D. R. Nayak, and M. A. Zaveri, "A comprehensive analysis of dermoscopy images for melanoma detection via deep cnn features," *Biomedical Signal Processing and Control*, vol. 79, p. 104186, 2023.

[53] S.-L. Yi, S.-L. Qin, F.-R. She, and T.-W. Wang, "Red-cnn: The multi-classification network for pulmonary diseases," *Electronics*, vol. 11, no. 18, p. 2896, 2022.

[54] Y. R. Choi, S. H. Yoon, J. Kim, J. Y. Yoo, H. Kim, and K. N. Jin, "Chest radiography of tuberculosis: Determination of activity using deep learning algorithm." *Tuberculosis and Respiratory Diseases*, 2023.

[55] M. Nahiduzzaman, M. O. F. Goni, R. Hassan, M. R. Islam, M. K. Syfullah, S. M. Shahriar, M. S. Anower, M. Ahsan, J. Haider, and M. Kowalski, "Parallel cnn-elm: A multiclass classification of chest x-ray images to identify seventeen lung diseases including covid-19," *Expert Systems with Applications*, p. 120528, 2023.

[56] A. Iqbal, M. Usman, and Z. Ahmed, "Tuberculosis chest x-ray detection using cnn-based hybrid segmentation and classification approach," *Biomedical Signal Processing and Control*, vol. 84, p. 104667, 2023.

[57] I. A. Ahmed, E. M. Senan, H. S. A. Shatnawi, Z. M. Alkhraisha, and M. M. A. Al-Azzam, "Multi-techniques for analyzing x-ray images for early detection and differentiation of pneumonia and tuberculosis based on hybrid features," *Diagnostics*, vol. 13, no. 4, p. 814, 2023.

[58] A. Sultana, M. Nahiduzzaman, S. C. Bakchy, S. M. Shahriar, H. I. Peyal, M. E. Chowdhury, A. Khandakar, M. Arselene Ayari, M. Ahsan, and J. Haider, "A real time method for distinguishing covid-19 utilizing 2d-cnn and transfer learning," *Sensors*, vol. 23, no. 9, p. 4458, 2023.

[59] Z. Zulkifli, R. A. Soeprihatini, S. Sfenrianto, Z. Wiyanti, P. Bintoro, F. Fitriana, S. Sukarni, N. A. Putri, and D. Y. A. Andini, "Expert system for diagnosis of lung disease from x-ray using cnn and svm," *International Journal of Artificial Intelligence Research*, vol. 6, no. 2, 2022.

[60] M. Yildirim and A. Cinar, "Classification with respect to colon adenocarcinoma and colon benign tissue of colon histopathological images with a new cnn model: Ma_colonnet," *International Journal of Imaging Systems and Technology*, vol. 32, no. 1, pp. 155–162, 2022.

[61] D. Albashish, "Ensemble of adapted convolutional neural networks (cnn) methods for classifying colon histopathological images," *PeerJ Computer Science*, vol. 8, p. e1031, 2022.

[62] M. M. Zafar, Z. Rauf, A. Sohail, A. R. Khan, M. Obaidullah, S. H. Khan, Y. S. Lee, and A. Khan, "Detection of tumour infiltrating lymphocytes in cd3 and cd8 stained histopathological images using a two-phase deep cnn," *Photodiagnosis and Photodynamic Therapy*, vol. 37, p. 102676, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1572100021004932

[63] S. Majumdar, P. Pramanik, and R. Sarkar, "Gamma function based ensemble of cnn models for breast cancer detection in histopathology images," *Expert Systems with Applications*, vol. 213, p. 119022, 2023.

[64] R. Sadagopan, S. Ravi, S. V. Adithya, and S. Vivekanandhan, "Polyeffnetv1: A cnn based colorectal polyp detection in colonoscopy images," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 237, no. 3, pp. 406–418, 2023.

[65] G. Yue, S. Li, R. Cong, T. Zhou, B. Lei, and T. Wang, "Attention-guided pyramid context network for polyp segmentation in colonoscopy images," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[66] J. Lewis, Y.-J. Cha, and J. Kim, "Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images," *Scientific Reports*, vol. 13, no. 1, p. 1183, 2023.

[67] A. Krenzer, M. Banck, K. Makowski, A. Hekalo, D. Fitting, J. Troya, B. Sudarevic, W. G. Zoller, A. Hann, and F. Puppe, "A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks," *Journal of Imaging*, vol. 9, no. 2, p. 26, 2023.

[68] M. Murugesan, R. M. Arieth, S. Balraj, and R. Nirmala, "Colon cancer stage detection in colonoscopy images using yolov3 msf deep learning architecture," *Biomedical Signal Processing and Control*, vol. 80, p. 104283, 2023.

[69] M. Souaidi, S. Lafraxo, Z. Kerkaou, M. El Ansari, and L. Koutti, "A multiscale polyp detection approach for gi tract images based on improved densenet and single-shot multibox detector," *Diagnostics*, vol. 13, no. 4, p. 733, 2023.

[70] I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun, "An efficient real-time colonic polyp detection with yolo algorithms trained by using negative samples and large datasets," *Computers in Biology and Medicine*, vol. 141, p. 105031, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521008258

[71] A. Alammari, A. R. Islam, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, "Classification of ulcerative colitis severity in colonoscopy videos using cnn," in *Proceedings of the 9th International Conference on Information Management and Engineering*, ser. ICIME 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 139–144. [Online]. Available: https://doi.org/10.1145/3149572.3149613

[72] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, and Y. Shin, "Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 180–193, 2020.

[73] A. Karaman, I. Pacal, A. Basturk, B. Akay, U. Nalbantoglu, S. Coskun, O. Sahin, and D. Karaboga, "Robust real-time polyp detection system design based on yolo algorithms by optimizing activation functions and hyper-parameters with artificial bee colony (abc)," *Expert Systems with Applications*, vol. 221, p. 119741, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423002427

[74] Y. Zheng, R. Zhang, R. Yu, Y. Jiang, T. W. C. Mak, S. H. Wong, J. Y. W. Lau, and C. C. Y. Poon, "Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 4142–4145.

[75] R. Kundu, H. Basak, A. Koilada, S. Chattopadhyay, S. Chakraborty, and N. Das, "Ensemble of cnn classifiers using sugeno fuzzy integral technique for cervical cytology image classification," *arXiv preprint arXiv:2108.09460*, 2021.

[76] Q. Zhao, S. Lyu, W. Bai, L. Cai, B. Liu, M. Wu, X. Sang, M. Yang, and L. Chen, "A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation," *arXiv preprint arXiv:2207.06799*, 2022.

[77] D. Wang, W. Dai, D. Tang, Y. Liang, J. Ouyang, H. Wang, and Y. Peng, "Deep learning approach for bubble segmentation from hysteroscopic images," *Medical & Biological Engineering & Computing*, vol. 60, no. 6, pp. 1613–1626, 2022.

[78] J. Song, S. Im, S. H. Lee, and H.-J. Jang, "Deep learning-based classification of uterine cervical and endometrial cancer subtypes from whole-slide histopathology images," *Diagnostics*, vol. 12, no. 11, p. 2623, 2022.

[79] K. Dilna, J. Anitha, A. Angelopoulou, E. Kapetanios, T. Chaussalet, and D. J. Hemanth, "Classification of uterine fibroids in ultrasound images using deep learning model," in *Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part III*. Springer, 2022, pp. 50–56.

[80] Y. Zhang, Z. Wang, J. Zhang, C. Wang, Y. Wang, H. Chen, L. Shan, J. Huo, J. Gu, and X. Ma, "Deep learning model for classifying endometrial lesions," *Journal of Translational Medicine*, vol. 19, pp. 1–13, 2021.

[81] P. Török and B. Harangi, "Digital image analysis with fully connected convolutional neural network to facilitate hysteroscopic fibroid resection," *Gynecologic and obstetric investigation*, vol. 83, no. 6, pp. 615–619, 2018.

[82] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging mobilenet and transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5270–5279.

[83] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30 392–30 400, 2021.

[84] Q. Jia and H. Shu, "Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II*. Springer, 2022, pp. 3–14.

[85] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*. Springer, 2021, pp. 171–180.

[86] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[87] J. C. X. N. W. Qin, R.; Gan, "Research progress of hysteroscopic surgery in the treatment of uterine lesions," *Med. Rev.*, vol. 26, p. 3282–3286, 2020.

[88] S. Wang and Z. Qunying, "Analysis of reproductive prognosis of patients with different types of submucous myoma treated by hysteroscopic electrotomy," *J. Wannan Med. Coll*, vol. 38, pp. 260–263, 2019.

[89] K. Ramamurthy, T. T. George, Y. Shah, and P. Sasidhar, "A novel multi-feature fusion method for classification of gastrointestinal diseases using endoscopy images," *Diagnostics*, vol. 12, no. 10, p. 2316, 2022.

[90] P. Muruganantham and S. M. Balakrishnan, "Attention aware deep learning model for wireless capsule endoscopy lesion classification and localization," *Journal of Medical and Biological Engineering*, vol. 42, no. 2, pp. 157–168, 2022.

[91] D. Jha, S. Ali, N. K. Tomar, H. D. Johansen, D. Johansen, J. Rittscher, M. A. Riegler, and P. Halvorsen, "Real-time polyp detection, localization and segmentation in colonoscopy using deep learning," *Ieee Access*, vol. 9, pp. 40 496–40 510, 2021.

[92] S. Durak, B. Bayram, T. Bakırman, M. Erkut, M. Doğan, M. Gürtürk, and B. Akpınar, "Deep neural network approaches for detecting gastric polyps in endoscopic images," *Medical & Biological Engineering & Computing*, vol. 59, pp. 1563–1574, 2021.

[93] A. Yamada, R. Niikura, K. Otani, T. Aoki, and K. Koike, "Automatic detection of colorectal neoplasia in wireless colon capsule endoscopic images using a deep convolutional neural network," *Endoscopy*, vol. 53, no. 08, pp. 832–836, 2021.

[94] S. M. Zhang, Y. J. Wang, and S. T. Zhang, "Accuracy of artificial intelligence-assisted detection of esophageal cancer and neoplasms on endoscopic images: a systematic review and meta-analysis," *Journal of Digestive Diseases*, vol. 22, no. 6, pp. 318–328, 2021.

[95] E. Hodneland, J. A. Dybvik, K. S. Wagner-Larsen, V. Šoltészová, A. Z. Munthe-Kaas, K. E. Fasmer, C. Krakstad, A. Lundervold, A. S. Lundervold, Ø. Salvesen *et al.*, "Automated segmentation of endometrial cancer on mr images using deep learning," *Scientific reports*, vol. 11, no. 1, pp. 1–8, 2021.

[96] Y. Kurata, M. Nishio, Y. Moribata, A. Kido, Y. Himoto, S. Otani, K. Fujimoto, M. Yakami, S. Minamiguchi, M. Mandai *et al.*, "Automatic segmentation of uterine endometrial cancer on multi-sequence mri using a convolutional neural network," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[97] Y. Takahashi, K. Sone, K. Noda, K. Yoshida, Y. Toyohara, K. Kato, F. Inoue, A. Kukita, A. Taguchi, H. Nishida *et al.*, "Automated system for diagnosing endometrial cancer by adopting deep-learning technology in hysteroscopy," *PLoS One*, vol. 16, no. 3, p. e0248526, 2021.

[98] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[99] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[100] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[101] H.-C. Dong, H.-K. Dong, M.-H. Yu, Y.-H. Lin, and C.-C. Chang, "Using deep learning with convolutional neural network approach to identify the invasion depth of endometrial cancer in myometrium using mr images: a pilot study," *International journal of environmental research and public health*, vol. 17, no. 16, p. 5993, 2020.

[102] B. Nikmard, N. Movahhedinia, and M. R. Khayyambashi, "Congestion avoidance by dynamically cache placement method in named data networking," *The Journal of Supercomputing*, pp. 1–27, 2022.

[103] X. Wang, N. Bao, X. Xin, J. Tan, H. Li, S. Zhou, and H. Liu, "Automatic evaluation of endometrial receptivity in three-dimensional transvaginal ultrasound images based on 3d u-net segmentation," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 8, p. 4095, 2022.

[104] Z. Ahmed, M. S. Kareem, H. A. Khan, F. H. Jaskani, Z. Saman, and B. Mughal, "Detection of uterine fibroids in medical images using deep neural networks."

[105] S. Sundar and S. Sumathy, "Transfer learning approach in deep neural networks for uterine fibroid detection," *International Journal of Computational Science and Engineering*, vol. 25, no. 1, pp. 52–63, 2022.

[106] Y.-H. Luo, I. L. Xi, R. Wang, H. O. Abdallah, J. Wu, A. Z. Vance, K. Chang, M. Kohi, L. Jones, S. Reddy *et al.*, "Deep learning based on mr imaging for predicting outcome of uterine fibroid embolization," *Journal of Vascular and Interventional Radiology*, vol. 31, no. 6, pp. 1010–1017, 2020.

[107] C.-m. TANG, D. LIU, and X. YU, "Mri image segmentation system of uterine fibroids based on ar-unet network," *American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)*, vol. 71, no. 1, pp. 1–10, 2020.

[108] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[109] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[110] N. Fatima, G. Bharti, and N. Mishra, "Role of hysteroscopy in the diagnosis of endometrial cancer."

[111] A. Zhao, X. Du, S. Yuan, W. Shen, X. Zhu, and W. Wang, "Automated detection of endometrial polyps from hysteroscopic videos using deep learning," *Diagnostics*, vol. 13, no. 8, 2023.

[112] D. Djokovic and A. Drizi, "Diagnostic hysteroscopy: patient assessment and preparation."

[113] Y. Zhang, C. Gong, L. Zheng, X. Li, and X. Yang, "Deep learning for intelligent recognition and prediction of endometrial cancer," *Journal of Healthcare Engineering*, vol. 2021, 2021.

[114] A. Urushibara, T. Saida, K. Mori, T. Ishiguro, K. Inoue, T. Masumoto, T. Satoh, and T. Nakajima, "The efficacy of deep learning models in the diagnosis of endometrial cancer using mri: a comparison with radiologists," *BMC Medical Imaging*, vol. 22, no. 1, pp. 1–14, 2022.

[115] Y. Zhang, Z. Wang, J. Zhang, C. Wang, Y. Wang, H. Chen, L. Shan, J. Huo, J. Gu, and X. Ma, "Deep learning model for classifying endometrial lesions," *Journal of Translational Medicine*, vol. 19, pp. 1–13, 2021.

[116] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[117] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[118] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[119] A. Goyal, A. Bochkovskiy, J. Deng, and V. Koltun, "Non-deep networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6789–6801, 2022.

[120] D. Zhou, F. Tian, X. Tian, L. Sun, X. Huang, F. Zhao, N. Zhou, Z. Chen, Q. Zhang, M. Yang, Y. Yang, X. Guo, Z. Li, J. Liu, J. Wang, J. Wang, B. Wang, G. Zhang, B. Sun, and X. Li, "Diagnostic evaluation of a deep learning model for optical diagnosis of colorectal cancer," *Nature Communications*, vol. 11, 06 2020.