Biomedical Data Science
2026, VOL. 6, NO. 1, 116-121
DOI: 10.47297/wspbdsWSP2752-630526.20260601

Semantic Segmentation of Malignant Tumors in Breast Ultrasound Images Based on Lightweight U-Net

Yusen Li

Tianjin University of Traditional Chinese Medicine, Medical Information Engineering, Tianjin, 301617, China

ABSTRACT

To address the challenges of high noise, blurred edges in breast ultrasound images, and the large parameter size and slow inference speed of traditional U-Net models, which fail to meet the demands of real-time clinical auxiliary diagnosis, a lightweight U-Net-based semantic segmentation method for malignant tumors in breast ultrasound images with multi-scale supervision is proposed. This algorithm adopts the classical U-Net architecture as its foundation, introduces multi-scale supervision strategies and channel attention mechanisms to enhance tumor feature extraction capabilities, while simultaneously employing structured pruning and quantized attention training (QAT) to achieve extreme model lightweighting. Experimental results demonstrate that the lightweight model maintains segmentation accuracy (Dice coefficient reaching 0.818) while reducing parameter size by approximately 76.8% and improving inference speed by about 77.7%. This approach provides efficient technical support for intelligent auxiliary diagnosis of malignant breast tumors, making it suitable for low-power, real-time diagnostic scenarios in primary healthcare institutions.

KEYWORDS

Deep learning; Semantic segmentation; Breast ultrasound images; U-Net model; model lightweighting; Pruning; Quantization-perception training

1 Introduction

Breast cancer is the most prevalent malignant tumor among women globally, representing the leading cancer threat to women's health and life. Its incidence and mortality rates rank among the highest for female malignancies, making it a global public health issue. According to 2020 GLOBOCAN statistics, there are approximately 2.3 million new breast cancer cases worldwide each year, accounting for about 24.5% of all new female malignancies. China's annual incidence rate reaches 39.1 per 100,000, with a significant trend of rising incidence and younger patient demographics, indicating an increasingly severe situation [1]. The harm of breast cancer manifests in multiple aspects, including physiological, psychological, and socioeconomic impacts. Early symptoms are highly insidious, often presenting as painless breast lumps that are easily overlooked. Once the disease progresses to intermediate or advanced stages, cancer cells can metastasize through lymphatic and blood systems, causing severe physical suffering to patients. In terms of survival rates, data from 17 cancer registries between 2003 and 2005 showed that the 1-year, 3-year, and 5-year observed survival rates for breast cancer patients were 90.5%,80.0%, and 72.7%, respectively, with a 5-year relative survival rate of 73.0% (95% CI: 0.712 – 0.749) [3]. For this reason, early screening, diagnosis, and treatment of breast cancer are crucial for reducing mortality and improving prognosis. Ultrasound examination is one of the conventional diagnostic methods for breast tumors. From grayscale ultrasound and real-time ultrasound in the 1980s to the application of color Doppler ultrasound, three-dimensional ultrasound, acoustic contrast imaging, ultrasonic elastography, and ultrasound light scattering breast imaging since the 1990s, the diagnostic and differential diagnostic rates of breast cancer through ultrasound have gradually improved, and its clinical guiding significance for breast cancer diagnosis and treatment has become increasingly prominent [4-5].

However, breast ultrasound images inherently suffer from issues such as excessive speckle noise, low tissue contrast, blurred tumor margins, and uneven gray distribution. Manual interpretation heavily relies on physicians' clinical experience, presenting drawbacks including time-consuming processes, strong subjectivity, and high rates of missed or misdiagnoses [6]. While deep learning-based intelligent medical image analysis, such as the U-Net model specifically designed for biomedical image segmentation, has demonstrated excellent performance in semantic segmentation tasks and achieves high-precision pixel-level segmentation on small sample datasets [7], traditional U-Net models face challenges like excessive parameter size, high computational complexity, and significant inference latency. These limitations hinder their deployment on low-power hardware platforms in mobile medical devices or primary healthcare institutions, failing to meet the demands of real-time clinical diagnostic assistance. Current domestic and international research on breast ultrasound image segmentation primarily focuses on improving segmentation accuracy, with insufficient attention to model lightweighting and practical deployment performance. Systematic exploration of channel redundancy compression and low-bit quantization in attention mechanisms remains lacking [8]. To address these issues, this study builds upon the U-Net model by constructing a multi-scale supervised architecture to enhance tumor feature extraction capabilities. Simultaneously, structured pruning is employed to eliminate redundant channels, and quantization-perception training is applied to reduce computational precision, achieving extreme model lightweighting. This approach improves inference speed while maintaining segmentation accuracy, providing a feasible solution for real-time intelligent auxiliary diagnosis and deployment of malignant breast ultrasound tumors with reduced implementation complexity.

2 Relevant Theories and Technical Foundations

2.1 U-Net model

U-Net is a classic encoder-decoder architecture fully convolutional neural network specifically designed for biomedical image segmentation. The model adopts a U-shaped structure divided into two components: the encoder and the decoder. The encoder employs a downsampling operation combining convolution and max pooling to progressively extract deep semantic features of the image, achieving feature dimensionality reduction. The decoder utilizes an upsampling operation through transposed convolution to restore image resolution, while also incorporating shallow feature representations from corresponding encoder layers via skip connections, effectively compensating for information loss during upsampling. The U-Net model demonstrates outstanding performance in small-sample medical image segmentation tasks without requiring massive labeled data, establishing itself as a benchmark model for medical image semantic segmentation. However, the presence of numerous redundant channels in both the encoder and decoder leads to large parameter sizes and slow inference speeds [9].

2.2 Model Lightweight Technology

Model lightweighting is a technical approach that reduces parameter quantity and computational complexity while maintaining model performance, primarily through pruning redundant parameters and reducing computational precision. It mainly includes two core methods: pruning and quantization [10]. Structured pruning directly decreases model parameters and computational load by trimming redundant channels, convolution kernels, and other structured units in neural networks, enabling deployment without specialized hardware [11]. Quantization-aware training (QAT) simulates errors from low-precision quantization (e. g., INT8) by inserting quantization pseudo-nodes during training, allowing models to adapt to quantized computation. This method minimizes computational precision while preserving model performance, effectively reducing memory usage and computation time [12]. This paper adopts a "prune first, then quantify" strategy to avoid error accumulation caused by quantized redundant parameters, thereby enhancing lightweighting effects.

2.3 Multi-scale Supervision and Attention Mechanism

Multi-scale supervision, by extracting features at different scales of images and conducting supervised training, captures multi-scale morphological features of targets, thereby enhancing the model's segmentation capability for tumors of varying sizes and shapes. The attention mechanism automatically learns the weights of feature channels, enabling the model to focus on feature information useful for segmentation tasks while suppressing interference from irrelevant background information, significantly improving the model's feature extraction and segmentation accuracy [13]. Combining multi-scale supervision with the attention mechanism can effectively address the segmentation challenges of blurred tumor margins and variable morphologies in breast ultrasound images.

3 Methodology of This Paper

The lightweight U-Net-based malignant tumor segmentation model for breast ultrasound images proposed in this paper is constructed on the foundation of classical U-Net. Through a multi-scale supervised architecture and channel attention mechanism, it enhances tumor feature extraction capabilities. Combined with structured pruning and quantization-aware training, the model achieves extreme lightweight design, with an overall architecture that balances segmentation accuracy and inference efficiency.

3.1 Construction of Multi-scale Supervision Architecture

This study presents a multi-scale supervision architecture specifically designed for thyroid nodule segmentation in two-dimensional ultrasound images, with its overall architecture illustrated in Figure 1. The model is built upon the U-Net architecture and consists of three core components: encoder, decoder, and multi-scale supervision. The encoder generates four semantic feature maps at different scales through four groups of downsampled convolutional operations, with the fourth feature map fed into an improved Atrous Spatial Pyramid Pooling (ASPP) module to extract more representative feature information. The decoder recovers detailed information from the semantic feature map via upsampled and skip connections, followed by four convolutional layers to transform features into five sizes: $16\times16\times1$, $32\times32\times1$, $64\times64\times1$, $128\times128\times1$, and $256\times256\times1$. Each convolutional module incorporates an improved attention mechanism to achieve deep integration of feature information across adjacent layers. In the encoder, a maximum pooling layer with a kernel size of 3×3, stride of 2 pixels, and padding of 1 pixel is connected after the convolutional layers to perform downsampling. In the decoder, an upsampled layer with a stride of 2 pixels is connected after the convolutional layers to restore resolution. Finally, the softmax layer generates probability distributions of the decoder's output depth features across two categories of labels.

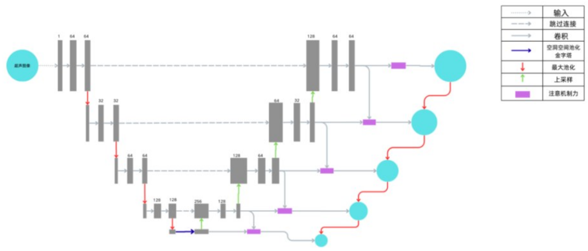

Figure 1 Overall Architecture of the Model

3.2 Channel Attention Mechanism Embedding

The channel attention module is embedded into the multi-scale supervision mechanism of the decoder to achieve precise extraction of effective features. After processing by the decoder, the feature maps generate feature maps X and Y at different scales. Feature map X undergoes 4-fold global maximum pooling, while feature map Y undergoes global average pooling. The pooled features are then transformed nonlinearly through two fully connected layers with ReLu activation function. The transformed features are subsequently fused by channel-wise summation, and the resulting feature channel weights are generated via the Sigmoid function. These weights are then multiplied channel-wise with the original feature map Y to produce a feature map with attention weights. This operation enables the model to focus on effective feature information of breast tumors, suppresses interference from noise and background in ultrasound images, and improves segmentation accuracy [13].

3.3 Structured Pruning Strategy

To address the issues of excessive parameters and redundant channels in traditional U-Net, we implemented structural pruning to eliminate redundant channels, achieving initial model lightweighting. The pruning process adhered to the core principle of retaining the decoder while pruning the encoder [14], utilizing the PAT pruning model for channel pruning.

Encoder pruning: The core function of the encoder is to extract deep semantic features, while its shallow channels primarily capture basic texture features of images, which have minimal impact on the details of breast tumor segmentation. Therefore, the focus is on pruning redundant shallow channels in the encoder, with a pruning ratio of 60%~70%, to reduce parameter quantity and computational load.

Decoder retention: The decoder is responsible for restoring image resolution and achieving fine segmentation. It requires sufficient channels to ensure feature fusion and detail restoration. Therefore, only the most obvious redundant channels (less than 10% of the total channels) are removed from the decoder to avoid a significant drop in segmentation accuracy.

After pruning, the model undergoes fine-tuning with a learning rate set at 1/10 of the initial training rate and 30 epochs to recover performance loss caused by channel pruning, ensuring segmentation accuracy.

3.4 Optimization of Quantitative Perception Training

Building upon structured pruning, we implement a hybrid strategy combining Quantitative Attention Training (QAT) with per-channel quantization. The QuantU-Net quantization model is employed to achieve model lightweighting while preserving maximum segmentation accuracy. The core quantization strategy is as follows:

(1) INT8 quantization pseudo-node is inserted into the training process of the model after pruning and fine-tuning to simulate the error caused by low precision quantization, so that the model can adapt to the quantization calculation mode during training and reduce the impact of quantization on the model performance.

(2) The quantization parameters of each channel are calculated separately, which is better than the tensor quantization, and the feature expression ability of the convolution layer is better preserved.

(3) The attention module employs FP16 high-precision quantization strategy to prevent the attention mechanism from failing due to low-precision quantization, ensuring the model's ability to focus on effective features.

The training parameters of quantization-aware training are consistent with the baseline model, ensuring the model fully learns the quantization error compensation strategy.

4 Experimental Design and Result Analysis

4.1 Experimental Data Set

This experiment utilized the Kaggle public breast ultrasound dataset, which comprises clinical breast ultrasound imaging data collected in 2018. The dataset covers female patients aged 25–75 years, with a total of 600 patients and 780 PNG images included. The average image size is 500×500 pixels. The images are classified into three categories—normal, benign, and malignant—based on clinical diagnosis. Each sample includes both the original ultrasound image and a physician-annotated true segmentation mask image, providing reliable clinical data support for breast ultrasound image segmentation and benign/malignant differentiation. Additionally, all experiments in this paper focused on segmentation tasks, and the proposed method's effectiveness was validated using evaluation metrics such as recall, precision, Dice coefficient, and accuracy.

4.2 Data Preprocessing

(1) Dimensional normalization: All images are uniformly resized to 256×256 pixels to match the model input dimensions while reducing computational load.

(2) Pixel value normalization: Map image pixel values from [0,255] to [0,1] range to eliminate the impact of pixel value scale differences on model training;

(3) ROI extraction: Based on the malignant tumor annotation mask, calculate the lesion center coordinates, and crop a 256×256 pixel region of interest (ROI) with the center as the origin, eliminating irrelevant background information to enhance tumor features;

(4) Data augmentation: To address the issues of limited sample quantity and uneven distribution of abnormal samples (malignant tumors) in medical image datasets, a combined data augmentation method was employed, including "random horizontal/vertical flipping (probability 0.5), random rotation (-30° ~30°), brightness adjustment (±0.2), and contrast adjustment (±0.2)". This approach expanded the dataset to 3,120 images, which were then divided into training set, validation set, and test set in a 7: 2: 1 ratio, thereby enhancing the model's generalization ability and accuracy in detecting abnormal samples.

4.3 Experimental Hardware and Software Environment

Hardware configuration: The experiment employed a single GPU for both training and inference, with the entire process running on an Nvidia Titan XP 12GB (CUDA Version 11.7) graphics card and an Intel® Xeon® Bronze 3106 CPU@ 1.70 GHz × 8.

Software environment: The experiment was built using Python deep learning frameworks, with the following specific software versions: Programming language: Python 3.9; Deep learning frameworks: PyTorch 1.13.1 and Torchvision 0.14.1; Image processing libraries: OpenCV 4.7.0, PIL 9.4.0, and NumPy 1.24.3.

4.4 Experimental Parameter Settings

4.4.1 Baseline Model Training Parameters

Using the classic U-Net as the baseline model, the training parameters were set with a batch size of 8, an initial learning rate of 0.01, and a learning rate decay strategy of StepLR. After training for 20 and 100 epochs, the learning rate was reduced by a factor of 10 to ensure full convergence of the model. The optimizer was Adam, with Beta1=0.9, Beta2=0.999, and weight decay rate = 1×10^{-8}. The loss function was the Dice loss plus the weighted sum of binary cross-entropy (BCE) loss (both with weight 1), evaluating the similarity between segmentation results and true annotations from both global and local perspectives. The training epochs were set to 150. Additionally, the training was stopped after 10 consecutive epochs without improvement in the Dice coefficient of the validation set to avoid premature termination due to overfitting.

4.4.2 Pruning and Quantitative Training Parameters

Structured pruning: Perform PAT pruning using the baseline model with a threshold of 0.05. After pruning redundant channels, fine-tune for 30 epochs at a learning rate of 0.001, while maintaining all other parameters consistent with the baseline model.

Quantized attention training: The QAT training is performed on the pruned and fine-tuned model by inserting INT8 quantization pseudo-nodes. The training epoch count is 50, with the same learning rate and optimizer parameters as the baseline model. Quantization errors are monitored in real-time during training.

4.5 Experimental Evaluation Indicators

The model is comprehensively evaluated from two dimensions: segmentation accuracy and model lightweight performance. Both assessments utilize test set data to ensure objective results.

4.5.1 Splitting Precision Index

The classic evaluation metrics for medical image semantic segmentation tasks are all pixel-level assessments, where higher metric values closer to 1 indicate greater segmentation accuracy.

Dice coefficient (DSC): A core metric for evaluating medical image segmentation, it measures the overlap between segmented results and true labels. The formula is: $LDice = 1 - \sum g(x, y)^2 + \sum p(x, y)^2 + \varepsilon^2 \sum g(x, y) \times p(x, y)$, where $g(x, y)$ denotes the true label of pixel (x, y), $p(x, y)$ is the model's prediction, and ε is a constant (set to 1e-8) to prevent the denominator from becoming zero.

Precision: The proportion of pixels predicted as tumors by the model that are actually tumors, reflecting the accuracy of segmentation.

Recall: A metric that measures the proportion of true tumor pixels correctly predicted by the model, reflecting the completeness of segmentation.

Pixel accuracy (Accuracy): The percentage of pixels predicted correctly by the model.

4.5.2 Lightweight Performance Indicators

Model parameters (Params): Measured in megabytes (M) to reflect memory usage, calculated using the torchsummary library.

Single-image inference time: measured in milliseconds (ms). The model's real-time inference capability is evaluated by calculating the average inference time across 100 randomly selected images from the test set.

Parameter compression ratio: Measures the reduction in model parameters after pruning. The formula is: Compression ratio = (Lightweight model parameters / Baseline model parameters)

Inference speed improvement rate: Measures the percentage increase in inference speed of the lightweight model. The calculation formula is: Improvement rate = (Inference time of lightweight model-Inference time of baseline model) / Inference time of baseline model

4.6 Experimental Results and Analysis

4.6.1 Comparison of Model Performance Results

The proposed multi-scale supervised lightweight U-Net model is compared with the classical U-Net baseline model and the unlightweight U-Net model (without multi-scale supervision). The experimental results are shown in Table 1.

Table 1 Comparison of Performance Among Different Models

types of models	Dice	Precision	recall	pixel accuracy	Parameter (M)	Inference time (ms)
classics U-Net	0.823	0.815	0.831	0.932	31.0	86.5
Multi-scale Supervised U-Net	0.845	0.839	0.852	0.946	35.2	98.7
Lightweight U-Net in this paper	0.818	0.812	0.825	0.928	7.2	19.3

4.6.2 Result Analysis

The Dice coefficient of the multi-scale supervised U-Net model reached 0.845, representing a 2.2% improvement over the classical U-Net. This indicates that the multi-scale supervision architecture and channel attention mechanism can effectively enhance the model's tumor feature extraction capability, addressing issues such as blurred tumor margins and morphological variability in breast ultrasound images, thereby improving segmentation accuracy. The Dice coefficient of the lightweight U-Net model in this study was 0.818, showing only a 0.5% decrease compared to the classical U-Net. The precision, recall, and pixel accuracy all experienced only minor declines, remaining above 0.8. This demonstrates that the pruning and quantization strategies employed in this study significantly reduced the model parameter size while maximizing segmentation accuracy, with performance meeting the clinical requirements for auxiliary diagnosis of malignant breast tumors in ultrasound imaging.

This lightweight U-Net model reduces parameter size from 31.0M to 7.2M, achieving a 76.8% compression rate that significantly decreases memory footprint for deployment on low-memory mobile medical devices. The inference time per image decreases from 86.5ms to 19.3ms, with a 77.7% speedup that resolves the high latency issue of traditional U-Net models, meeting real-time clinical diagnostic needs. Notably, the multi-scale supervised U-Net model, despite incorporating multi-scale supervision and attention modules, shows increased parameter size and inference time. This demonstrates that improving segmentation accuracy requires optimizing deployment performance through lightweight techniques to achieve clinical utility.

4.6.3 Ablation Test Results

To validate the effectiveness of each module in this model, we designed an ablation experiment to separately assess the impact of multi-scale supervision (MS), channel attention (CA), structured pruning (SP), and quantization-aware training (QAT) on model performance. The experimental results are presented in Table 2.

Table 2 Ablation Experiment Results

Model configuration	Dice coefficient	Parameter (M)	Inference time (ms)
foundation U-Net	0.823	31.0	86.5
Basic U-Net+MS	0.836	32.5	92.3
Basic U-Net+MS+CA	0.845	35.2	98.7
Basic U-Net+MS+CA+SP	0.839	8.5	25.6
Basic U-Net+MS+CA+SP+QAT	0.818	7.2	19.3

The ablation results demonstrate that the integration of multi-scale supervision and channel attention modules progressively improves the model's Dice coefficient, confirming their effectiveness in enhancing feature extraction and segmentation accuracy. The implementation of structured pruning reduces the model's parameter count from 35.2M to

8.5M while decreasing inference time to 25.6ms, with the Dice coefficient dropping only 0.6%—evidencing the pruning strategy's efficacy in achieving model lightweighting with minimal accuracy loss. Quantization-aware training further reduces parameters to 7.2M and inference time to 19.3ms, with the Dice coefficient decreasing by 2.1%, indicating a reasonable trade-off between quantization and performance.

5 Conclusion and Prospects

This study addresses the limitations of traditional U-Net models in malignant tumor segmentation for breast ultrasound images, which are characterized by excessive parameter quantities, slow inference speed, and inadequate performance for real-time clinical deployment. A lightweight U-Net segmentation model with multi-scale supervision is proposed, which enhances tumor feature extraction capabilities through a multi-scale supervision architecture and channel attention mechanism, while achieving extreme model lightweighting via structured pruning and quantization-aware training. Experimental results demonstrate that multi-scale supervision and channel attention mechanism significantly improve segmentation accuracy, with a Dice coefficient reaching 0.845, representing a 2.2% improvement over classical U-Net. The pruning and quantization strategy implemented in this paper achieves efficient model lightweighting, compressing parameters by 76.8% and improving inference speed by 77.7%, with only marginal loss in segmentation accuracy (Dice coefficient 0.818), meeting the precision requirements for clinical auxiliary diagnosis. This model balances segmentation accuracy and inference efficiency, substantially reducing memory consumption and inference latency, making it suitable for deployment on low-power hardware platforms such as mobile medical devices or primary healthcare institutions. It provides an efficient and feasible technical solution for real-time intelligent auxiliary diagnosis of malignant tumors in breast ultrasound.

The lightweight U-Net model proposed in this study achieves a balance between segmentation accuracy and inference efficiency, yet there remains room for further improvement. Subsequent research will focus on end-to-end system development, aiming to deploy this lightweight model onto mobile terminals such as smartphones and portable ultrasound devices. This will facilitate the development of an end-to-end intelligent auxiliary diagnostic system for breast ultrasound, integrating ultrasound image acquisition, malignant tumor segmentation, and benign-malignant differentiation. The goal is to advance the clinical application of artificial intelligence technology in primary breast disease screening. Additionally, breast ultrasound datasets collected from different hospitals and equipment will be compiled to conduct multicenter clinical validation of the model, optimizing its robustness to better adapt to real-world clinical scenarios.

References

[1] Chen Maoshan, Lü Qing. Key points interpretation of "Analysis of global incidence and mortality of breast cancer based on population registration data from 2000 to 2020" [J]. China Journal of Cardiothoracic Surgery, 2022,29(04):401-406.

[2] Zeng H, Zheng R, Guo Y, et al. Cancer survival in China, 2003.

[3] Chen Wanqing, Zheng Rongshou. Incidence, mortality and survival status of female breast cancer in China [J]. China Oncology Clinical, 2015, 42(13):668-674.

[4] Bao Mingwen, Xu Xiaohong. Current status and application of new technologies in ultrasound diagnosis of breast cancer [J]. Journal of Guangdong Medical University, 2009,27(02):204-205.

[5] He Xiaodong, Gu Suying. Current Status and Advances in Ultrasound Imaging for Breast Cancer[J]. Chinese Journal of General Practice, 2010,8 (02):231-233.

[6] Muhammad M, Zeebaree D, Brifcani AM, Muhammad M, Zeebaree D, Brifcani AM, et al. Region of interest segmentation based on clustering techniques for breast cancer ultrasound images:a review[J].J Appl Sci Technol Trends,2020,1(2):78-91.

[7] Zhao Kefu, Zhang Lei. Ultrasound image segmentation of thyroid nodules based on convolutional neural networks [J]. Modern Computer, 2021(15):54-60.

[8] Shi Jun, Wang Tiantong, Zhu Ziqi, et al. A review of medical image segmentation methods based on deep learning [J]. China Journal of Image and Graphics, 2025,30(06):2161-2186.

[9] Xu Guangxian, Feng Chun, Ma Fei. A Review of Medical Image Segmentation Based on U-Net [J]. Computer Science and Exploration, 2023,17 (08):1776-1792.

[10] Bi Pengcheng, Luo Jianxin, Chen Weiwei. Research on Lightweight Convolutional Neural Network Technology [J]. Computer Engineering and Applications, 2019,55(16):25-35.

[11] Lin Jingdong, Wu Xinyi, Chai Yi, Yin Hongpeng. Review on Structural Optimization of Convolutional Neural Networks [J]. Journal of Automation Chemistry, 2020,46(01):24-37.

[12] Yang Chun, Zhang Ruiyao, Huang Long, Ti Shutong, Lin Jinhui, Dong Zhiwei, Chen Songlu, Liu Yan, Yin Xucheng. A Review of Quantification Methods for Deep Neural Network Models [J]. Acta Engineering Scientiae, 2023,45(10):1613-1629.

[13] Zhou Xiaosong, Zhao Juanjuan. Multi-scale supervised U-Net for ultrasound image segmentation of thyroid nodules [J]. Journal of Taiyuan University of Technology, 2022,53(06):1134-1142.

[14] Jiang Xiaoyong, Li Zhongyi, Huang Langyue, et al. Review on Neural Network Pruning Techniques [J]. Acta Scientia Applicatae Sinica, 2022,40 (05):838-849.