

SwiF-YOLO: A Deep Learning Method for Lung Nodule Detection

Cheng Ren, Shouming Hou, Jianchao Hou and Yuteng Pang

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo Henan, 454003, China

Abstract: Lung cancer, a prevalent and lethal tumor globally, has a five-year survival rate of only 10%-16% for late-stage patients. However, early diagnosis and treatment can increase this rate to 52%. Lung nodules, as crucial indicators of early lung cancer, are challenging to detect due to their small size and similar features to other lung tissues. Therefore, developing an automatic detection method to improve the efficiency and accuracy of lung nodule detection is vital. This paper proposes a new method based on the YOLOx model, called SwiF-YOLO, to enhance the precision and efficiency of lung nodule detection. We introduced the Swin transformer to replace the main network of yolox-m, adopted the Adaptively Spatial Feature Fusion (ASFF) as the feature fusion method, and replaced the Intersection over Union (IOU) regression loss function with Generalized Intersection over Union (GIoU). These improvements aim to enhance the accuracy and efficiency of lung nodule detection, assisting doctors in diagnosing more accurately and quickly.

Keywords: Lung Nodule Detection; Object Detection; Medical Image Processing; Deep Learning.

1. Introduction

Lung cancer is the most common and deadly tumor disease worldwide [1]. Clinical studies have shown that the five-year survival rate for patients with advanced disease is only between 10% and 16%, but this can increase to 52% if early diagnosis and treatment are provided. As an important marker of early lung cancer, lung nodules mainly appear as localized, rounded lung shadows up to 3 cm in diameter on CT images. However, due to the tiny size of lung nodules and the similarity of their morphology, brightness, and other characteristics to blood vessels and other tissues in the lung parenchyma, physicians need to carefully consider and screen them one by one, a process that is inefficient, easily leads to fatigue, and increases the likelihood of misdiagnosis. Therefore, it is important to develop automated detection methods to help physicians improve the efficiency and accuracy of lung nodule detection [2].

Methods for lung nodule detection can be broadly categorized into two groups: traditional segmentation-based detection techniques and deep learning-based detection techniques [2]. Segmentation-based methods are mainly trained using manually extracted features, and the weaknesses of this method include cumbersome steps, low accuracy, and poor overall performance. With the rapid development of machine learning techniques, deep learning techniques have been widely used in target detection. Among them, works like Faster R-CNN and Mask R-CNN focus on two-stage detection algorithms based on candidate regions, the other is one-stage detection algorithms based on regression, such as the You Only Look Once (YOLO) algorithm and Single Shot MultiBox Detector (SSD). YOLOx, as a newer version of the YOLO algorithm, has already achieved notable successes in practical engineering applications.

In order to further improve the accuracy and efficiency of lung nodule detection, we propose a new method based on the YOLOx model called SwiF-YOLO. We make three significant improvements to YOLOx. First, we introduce the Swin transformer to replace the backbone network of yolox-m. The Swin transformer is a novel vision transformer model

that is effectively adapted to computer vision tasks by using displacement windows to construct hierarchical feature maps. Unlike traditional transformer models, Swin transformer's self-attention computation is limited to within a localized window, making the computational complexity linear, rather than quadratic, with the image size. This design not only improves the efficiency of the model, but also maintains a strong feature extraction capability. The innovation of the Swin transformer lies in its ability to capture both detailed and global information of an image at different levels, making it a powerful generalized backbone network for a variety of vision tasks.

Second, we adopt ASFF as a feature fusion method. ASFF is a feature fusion approach that can effectively represent feature maps at different levels and improve the model's ability to perceive targets at different scales. This approach improves feature scale invariance and enhances the importance of key layers.

Finally, we replace the regression loss function IOU with GIoU. GIoU is an improved method of calculating IoU that takes into account not only the intersection of the two bounding boxes, but also the area of the smallest bounding box that contains these two bounding boxes. All three improvements are aimed at improving the accuracy and efficiency of lung nodule detection to help physicians make more accurate and faster diagnoses.

2. Related Work

2.1. The Detection of Lung Nodules

Lung diseases, including chronic obstructive pulmonary disease (COPD), interstitial lung disease, pneumonia and lung cancer, can significantly impair lung function. Among them, lung cancer is one of the leading causes of illness and death worldwide, especially in China, where it is the second most common cancer in women and the leading cause of cancer-related death in men [3]. The key to improving the five-year survival rate of lung cancer is to detect, diagnose and treat it at an early stage [4]. Accurate identification of lung nodules is crucial as they are the initial indicators of lung cancer.

Detection of lung nodules by radiologists is a very cumbersome and time-consuming process, and is highly susceptible to missed or misdiagnosis. Computer aided diagnosis (CAD) systems can improve diagnostic efficiency and reduce diagnostic costs, and are of high research value. Computed tomography (CT) is becoming more and more popular in the detection and diagnosis of lung nodules due to its high sensitivity, fast acquisition speed, and cost-effectiveness [5].

In the early stages of diagnosing lung nodules, physicians rely on their previous experience to recognize specific symptoms of the disease. Similarly, researchers need to manually analyze features derived from medical images, which requires in-depth knowledge of the data and specialized clinical experience [6,7]. Due to rapid advances in computer vision and medical imaging, modern medicine has evolved from human analysis to automated processing using computer-aided systems and artificial intelligence techniques. Due to this change, early diagnosis of lung cancer is now more accurate and efficient. After a doctor obtains an image of the lungs through a CT scan, the CAD system begins to preprocess the CT image, and then the system uses target detection technology to identify lung nodules to help the doctor pinpoint the location of the lesion.

Traditional methods and machine learning algorithms were previously the main tools used to detect lung nodules. However, these methods have been gradually replaced by deep learning techniques. Deep learning algorithms can autonomously learn image features and achieve more accurate and efficient detection and classification by training on large amounts of data. Notably, lung nodule detection is often implemented using Convolutional Neural Network (CNN). These algorithms help physicians diagnose patients faster and more accurately, and improve the detection rate and diagnostic accuracy of lung cancer. Therefore, the application of deep learning algorithms to diagnose lung nodules has become a current research hotspot. Although deep learning techniques have made significant progress in the recognition of lung nodules, they still face challenges such as high data requirements, limited robustness of models, and detection of complex shapes and small targets.

2.2. Object Detection

Target detection is an important task in the field of computer vision, and its goal is to find specific classes of objects in an image or video and determine their location and size. The task of target detection can be mainly divided into four subtasks: Classification, which means that given an image to know what category of target is contained in it. Location, which gives the position of this target in the image. Detection, which locates the position of the target and determines the category of this target. Segmentation, which determines which target or scene each pixel belongs to. The core problems of target detection include: classification problem, which category the image (or a region) belongs to. Localization problem, where the target appears in the image. Size problem, the size of the target in the image. Shape problem, what kind of pose the target has in the image.

Deep learning is an approach to machine learning that uses multi-layer neural networks to learn and understand data by modeling the way the human brain works. In recent years, deep learning has made significant progress in the field of target detection. The development of deep learning has driven the advancement of target detection techniques. For example,

CNN's weight sharing and local connectivity greatly reduces the size of parameters and reduces the training complexity of the model, while the convolution operation preserves the spatial information of the image with translation invariance and certain rotation and scale invariance. These properties make CNNs well suited for target detection tasks.

The application of deep learning in target detection is mainly in the use of deep neural networks for feature extraction and classification. Deep neural networks can automatically learn and extract useful features from data, which makes them perform well in target detection tasks. Taking lung nodule detection as an example, deep learning can be used to detect and localize lung nodules from CT images. First, a CNN is used to extract features from the image. These features are then fed into a fully connected network that is used to predict the location (i.e., the parameters of the bounding box) and class (i.e., benign or malignant) of the lung nodule. This is typically done by predicting a bounding box, which is defined by four parameters: the coordinates of the center point (x, y), and the width and height of the bounding box (w, h).

Currently, there are two main technical routes for the application of deep learning in the field of target detection: the Anchor-based approach and the Anchor-free approach. Anchor-based methods mainly include two-stage and one-stage target detection algorithms.

(1) Two-stage target detection algorithm

A Region Proposal (RP), which is a pre-selected box that may contain the object to be detected, is preset first, and then the sample classification is calculated by the CNN. The process of this algorithm is feature extraction, generation of RP, and classification/regression localization. Common two-stage algorithms are R-CNN, Fast R-CNN [8], Faster R-CNN [9], Mask R-CNN [10], etc.

(2) Single-stage target detection algorithm

The feature values are extracted directly in the network to classify the target and localize it. The process of this algorithm is feature extraction and classification/regression localization. Single-stage target detection algorithms mainly include YOLO, SSD, RetinaNet, etc.

There are also some anchorless-based target detection algorithms such as CenterNet [11], CornerNet [12]. These algorithms try to avoid the use of anchor points by directly predicting the center or corner points of the target, thus simplifying the process of target detection.

Target detection is one of the core problems in the field of computer vision, where the task is to find out all the targets of interest in an image and determine their categories and locations. Target detection has always been a challenging problem due to the fact that different kinds of objects have different appearances, poses, and different degrees of occlusion, and imaging is the interference of lighting and other factors.

3. SwiF-YOLO

3.1. The Framework of SwiF-YOLO

The overall structure of the SwiF-YOLO model is shown in Figure 1. We make three significant improvements to YOLOx. First, we introduce Swin transformer to replace the backbone network of yolox-m. Swin transformer is a novel visual transformer model that is effectively adapted to computer vision tasks by using displacement windows to construct hierarchical feature maps. Unlike traditional

transformer models, Swin transformer's self-attention computation is limited to within a localized window, making the computational complexity linear, rather than quadratic, with the image size. This design not only improves the efficiency of the model, but also maintains a strong feature extraction capability. The innovation of the Swin transformer lies in its ability to capture both detailed and global information of an image at different levels, making it a powerful generalized backbone network for a variety of vision tasks.

Second, we adopt ASFF as a feature fusion method. ASFF is a feature fusion approach that can effectively represent

feature maps at different levels and improve the model's ability to perceive targets at different scales. This approach improves feature scale invariance and enhances the importance of key layers.

Finally, we replace the regression loss function IOU with GIoU. GIoU is an improved method of calculating IoU that takes into account not only the intersection of the two bounding boxes, but also the area of the smallest bounding box that contains these two bounding boxes. All three improvements are aimed at improving the accuracy and efficiency of lung nodule detection to help physicians make more accurate and faster diagnoses.

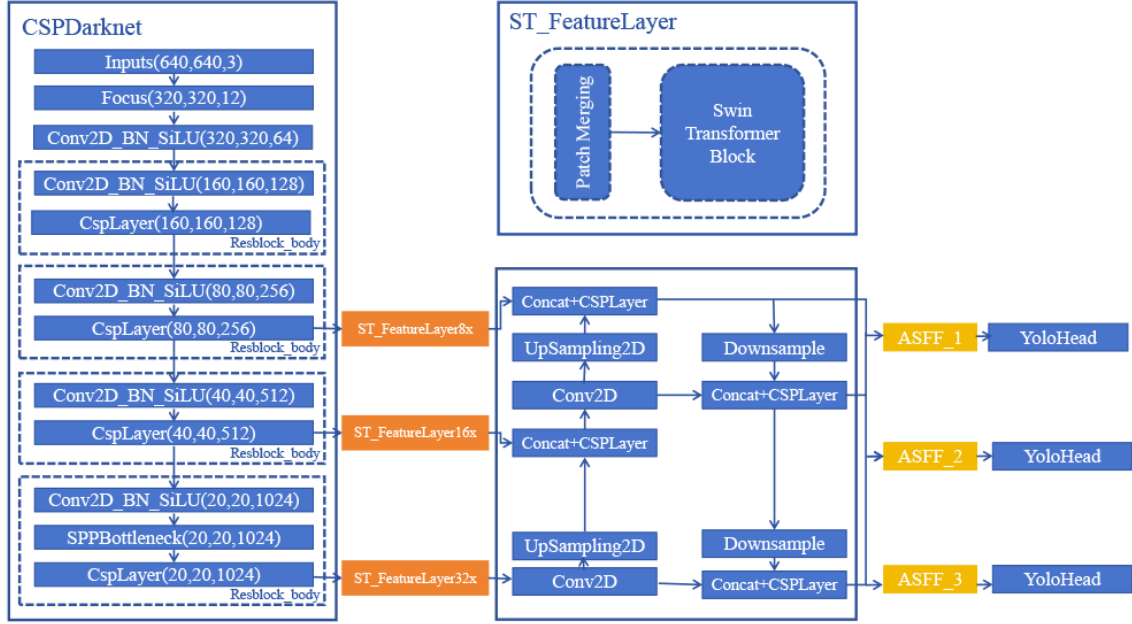


Fig 1. The structure of the SwiF-YOLO

3.2. Swin-transformer Network

Transformer was first designed by Vaswani et al [13] in the field of Natural Language Processing (NLP) and mainly consists of Multi-Head Attention (MHA) mechanism and

Position-wise Feed-Forward Networks. The original Transformer architecture is shown in Figure 2. Inspired by the Transformer, researchers started to improve the original Transformer in the field of computer vision, such as vision transformer (ViT) [14] and Swin Transformer [15].

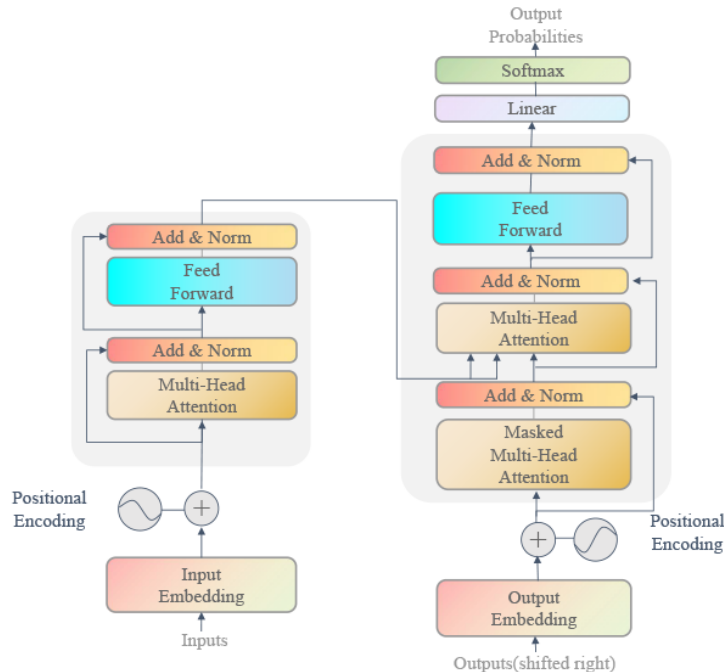


Fig 2. Transformer architecture

In NLP, the Transformer structure consists of an encoder and a decoder. However, in computer vision tasks, Swin Transformer uses only the encoder part of the Transformer. Therefore, the encoder part is mainly described next. The MHA module is a module in which multiple attention modules learn different aspects of attention in different subspaces. A matrix X consisting of vectors is first obtained and then mapped to different subspaces by learnable matrices W^Q , W^K , and W^V to obtain matrices query(Q), key(K), and value(V), respectively. In the encoder, Q , K , and V are equal, but it is worth noting that K and V in the second layer of the decoder come from the encoder and Q is the output of the first layer of the decoder. Q , K , and V can be written in the following form:

$$Q = XW^Q, K = XW^K, V = XW^V \quad (1)$$

where X is the matrix

$$W^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$$

The transpose of matrix Q and matrix K is used for dot product, and the similarity probability is obtained by scaling the softmax function by a factor of $\sqrt{d_k}$, where d_k is the dimension of matrix K . Finally, multiplication with V yields the attention weight matrix. The attention mechanism is executed in each header. The formula is as follows:

$$\text{Head}_i = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

Finally, the final feature representation is obtained by projecting through the learnable weight $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$. i denotes the number of heads.

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{Head}_1, \dots, \text{Head}_i)W^O \quad (3)$$

In the MHA mechanism, multiple subspaces can be formed with different heads focusing on different aspects of the information. Meanwhile, MHA enhances the stability and robustness of the network. The number of heads may vary in different works.

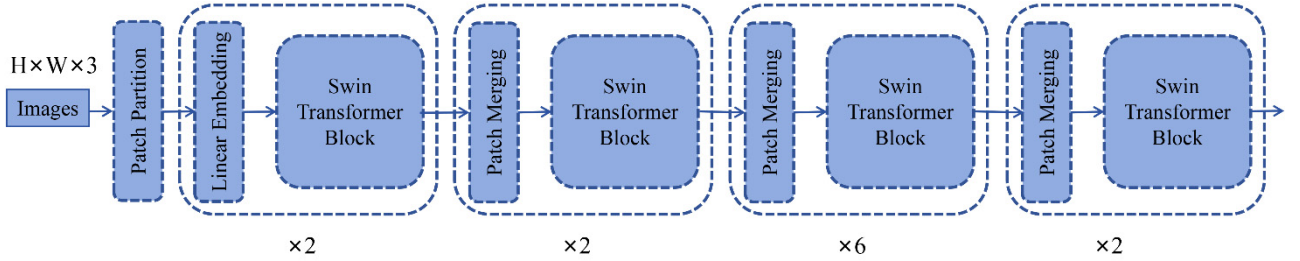


Fig 3. Swin Transformer architecture

Liu et al [15] proposed Swin Transformer, a ViT based on moving windows to improve computational efficiency. They designed hierarchical feature maps to obtain multi-resolution feature maps, which makes Swin Transformer a generalized backbone network for computer vision. The architecture of Swin Transformer is shown in Figure 3. This Swin Transformer block contains two sub-modules with window multi-head self attention (W-MSA) and shifted window multi-head self attention (SW-MSA) respectively replaced the MHA in ViT. The architecture of the Swin Transformer block is shown in Figure 4. In Swin Transformer, MSA and MLP are calculated as follows:

$$Z_n^{W-MSA} = W - \text{MSA}\left(\text{LN}\left(Z_{n-1}\right)\right) + Z_{n-1} \quad (4)$$

$$Z_n = \text{MLP}\left(\text{LN}\left(Z_n^{W-MSA}\right)\right) + Z_n^{W-MSA} \quad (5)$$

$$Z_{n+1}^{SW-MSA} = SW - \text{MSA}\left(\text{LN}\left(Z_n\right)\right) + Z_n \quad (6)$$

$$Z_{n+1} = \text{MLP}\left(\text{LN}\left(Z_{n+1}^{SW-MSA}\right)\right) + Z_{n+1}^{SW-MSA} \quad (7)$$

Since W-MSA and SW-MSA are designed in the Swin Transformer structure, the computation of attention is limited to each window instead of the whole image, reducing a large amount of computational complexity.

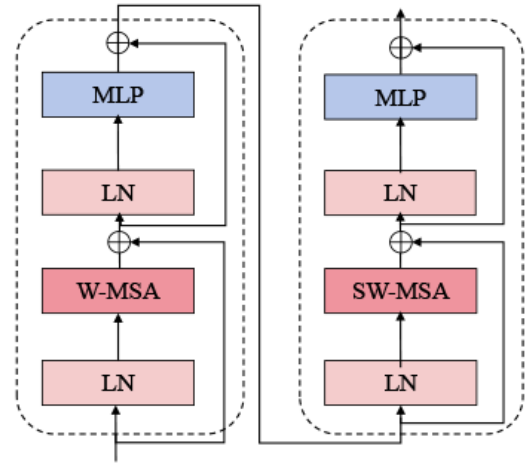


Fig 4. Swin Transformer block

3.3. Adaptively Spatial Feature Fusion

ASFF is a novel pyramid feature fusion strategy [16]. In CNNs, feature pyramids are an effective method to achieve target scale invariance. However, for single-order detectors based on feature pyramids, the inconsistency between features at different scales is one of the main limiting factors.

ASFF suppresses this inconsistency by learning how to spatially filter conflicting information, thus improving the scale invariance of the features and introducing a virtually unburdened inference overhead [16]. ASFF enables the network to directly learn how to spatially filter the features at other levels, thus retaining only the useful information for

combination.

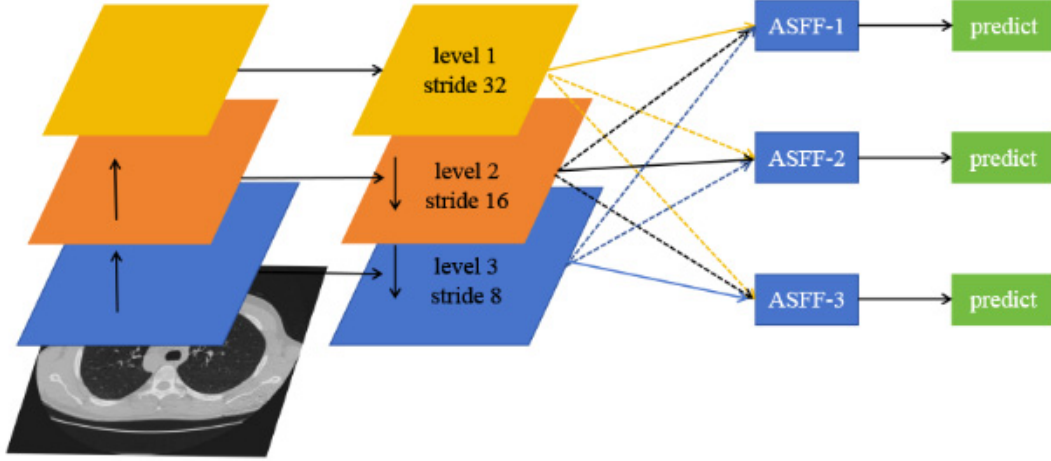


Fig 5. The structure of ASFF

The structure of ASFF is shown in Figure 5. For each level, the features at all other levels are tuned to the same shape and spatially fused based on the learned weight maps. The working principle of ASFF can be divided into two steps. First, the features at the other levels are tuned to the same resolution and simply integrated. Then, training is performed to find the best fusion. At each spatial location, features of different levels are adaptively fused together, e.g., if a location carries contradictory information, these features will be filtered out, and if the features at a location carry more distinguishing cues, these features will be enhanced.

The formula for ASFF can be expressed as:

$$ASFF - I_{ij} = \alpha_{ij}^l X^{1 \rightarrow l} + \beta_{ij}^l X^{2 \rightarrow l} + \gamma_{ij}^l X^{3 \rightarrow l} \quad (8)$$

Where $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$, and $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l \in [0, 1]$. These three parameters are obtained by 1×1 convolution.

ASFF has the following significant advantages: first, the process of searching for the optimal fusion is microscopic, which makes it easy to learn by backpropagation algorithms. Second, ASFF is independent of the type of the underlying network, and it can be widely applied to all single-order detectors with a feature pyramid structure. Finally, ASFF is implemented in a very concise manner with minimal additional computational cost. These advantages make ASFF promising for a wide range of applications in the field of target detection.

3.4. Improvement of Loss Function

YOLOX is a high-performance target detection model whose loss function is based on the combination of the cross-entropy loss function and the IoU loss function. The loss function of YOLOX is mainly computed in its YOLOXHead module. The YOLOXHead module is responsible for calculating the losses of each loss during the training process, including the confidence prediction, the category prediction and the bounding box prediction. Specifically, YOLOX uses a binary cross-entropy loss function to calculate the loss for confidence prediction, a cross-entropy loss function for category prediction, and an IOU loss for bounding box prediction. The YOLOX loss function can be expressed as follows:

$$L = L_{cls} + \lambda L_{reg} \quad (9)$$

Where L_{cls} denotes the categorization loss, L_{reg} denotes the regression loss, and λ is a hyperparameter used to balance the weights of the two losses. The specific classification loss and regression loss formulas are as follows:

$$L_{cls} = -\log(p_c) \quad (10)$$

$$L_{reg} = -\log(IoU) \quad (11)$$

Where p_c denotes the category probability of the predicted frame and IoU denotes the intersection and concurrency ratio of the predicted frame to the true frame. IoU can be expressed as follows:

$$IoU = \frac{A \cap B}{A \cup B} \quad (12)$$

The formula for the IoU loss function is as follows:

$$L_{IoU} = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (13)$$

Where A is the predicted bounding box and B is the real bounding box. From equation (13), we can see that the IoU loss has the following properties: (1) scale invariance: the IoU responds to the ratio between the intersection and concatenation of the two detection frames, and is therefore independent of the size of the detection frames; and (2) the IoU is a distance: this distance refers to a metric for evaluating the relationship between the two rectangular frames, which has all the properties of a distance, including symmetry, nonnegativity, homogeneity, and triangular inequality.

However, this loss function has some drawbacks. First, when there is no intersection between the predicted and real boxes, the IoU loss function has a value of 1, which makes the loss function impossible to optimize in this case. Second, the IoU loss function does not reflect the relative position and shape of the two frames, which may affect the performance of the model.

To address these issues, we consider replacing the original IoU loss function with the GIoU loss function, which is a metric that extends the traditional IoU by not only taking into account the intersection and concatenation of the two frames, but also introduces a new term to measure the closure of the two frames (i.e. the smallest rectangle containing the two boxes).

By using the GIoU loss function, we can evaluate the

similarity between two bounding boxes more accurately, thus improving the performance of the model. Especially in the case where there is no intersection of the two boxes, the GIoU loss function can provide more information to help optimize the model. Therefore, by replacing it with the GIoU loss function, we can effectively solve these shortcomings of the YOLOX loss function.

The GIoU loss function is an extension of the traditional IoU metric, which can more accurately assess the similarity between two bounding boxes. Its formula is as follows:

$$L_{GIoU} = 1 - IoU + \frac{|A_c - U|}{|A_c|} \quad (14)$$

Where A_c is the closure of the two regions and U corresponds to the area of the concatenation of the red and green rectangles.

4. Experimental Results

In this section, the experimental environment and the used

data set are first introduced, and then the model was trained on the public data set LUNA16. Next, an ablation experiment is performed on the LUNA16 data set to verify the effectiveness of the proposed detection method. Finally, the model was compared with other mainstream target detection models in terms of precision, recall and mAP.

4.1. Experimental Settings

This experiment configured Python for training and testing based on the Anaconda environment on the Windows 10 system, and used the PyTorch deep learning framework and CUDA for GPU acceleration to train and verify the feasibility of the model. PyTorch provides many extension libraries and interfaces for easy modular design, network construction and customization. In addition, PyTorch supports efficient parallel computing on CPU and GPU, and supports multi-machine distributed computing, which can greatly improve Training speed and model performance. The experimental configuration is shown in Table 1:

Table 1. The experimental configuration

Environment	Configuration
Programming Language	Python 3.8.18
Development tools	Pycharm 2023.2.5
Deep Learning Framework	Pytorch 2.1.1
Operating System	Windows 10x64
CPU	Core (TM) i5-13600KF
GPU	NVIDIA GeForce RTX 4070

4.2. Experimental Dataset

This article uses LUNA16 as the experimental data set. LUNA16 [17] is a high-quality lung nodule CT image data set launched in 2016. It is the most authoritative and representative data set in current pulmonary nodule detection research. The data set has a total of 888 3D lung CT images, 1186 pulmonary nodules and 36378 information annotated by 4 professional radiologists. The data set consists of four parts: original CT images, pulmonary nodule location annotation files, original CT lung region segmentation files, and diagnosis result files. This paper selects 70% of the pulmonary nodule samples in LUNA16 as the training set, 15% as the test set, and 15% as the validation set.

4.3. Model Training

This training data set uses the LUNA16 data set. Since the images in the LUNA16 data set are three-dimensional medical data, we first remove the two-dimensional slice graphics from the three-dimensional image, and then adjust the image to make the image resolution unified to 640×640.

Use stochastic gradient descent (SGD) as the optimizer for network training. In each iteration, SGD first randomly selects a portion of training samples (called a mini-batch), and then calculates the average loss and corresponding gradient of these samples. Then, SGD will update the parameters of the

model in the opposite direction of the gradient. This updated step size (called the learning rate) is an important hyperparameter. If the learning rate value is too high, it will be difficult for the model to find the global optimal solution. If the learning rate value is too small, the network will easily be limited to the local optimal solution. The advantage of SGD is that it can handle large-scale data sets efficiently because it only needs to process a portion of the training samples at a time. Furthermore, due to its stochastic nature, SGD is able to avoid falling into local minima of the loss function. However, SGD also has some disadvantages. For example, it is very sensitive to the choice of learning rate and other hyperparameters, and may take a long time to converge.

After comprehensive consideration, in this model training, the learning rate value was set to 0.01. The size of the batch during the training process will also affect the performance of the network model. If the batch is too large, it will help the model converge stably and reduce the model training time, but it will lead to a decrease in the generalization ability of the model. If the batch size is too small, the model will not converge easily and the training time will be increased, but it will have better generalization ability. After many experiments, the batch size was finally set to 4. The hyperparameter settings for this experiment are shown in Table 2.

Table 2. Experimental parameter configuration

Parameters	Batch size	Epoch	Learning rate	Weight decay
Values	4	300	0.01	0.0005

4.4. Ablation Experiments on SwiF-YOLO

In order to evaluate the contribution of each component in the model, we will remove each improvement point of the model one by one, including the Swin-Transformer backbone network, ASFF feature fusion, and GIoU loss function, and compare the performance changes of the model. Specifically, we will conduct an ablation experiment in the following steps:

(1) Swin-Transformer backbone network

We replace Swin-Transformer with the original YOLOX backbone network, keep other settings unchanged, and

observe changes in model performance.

(2) ASFF feature fusion

We will remove ASFF feature fusion, keep other settings unchanged, and observe the changes in model performance.

(3) GIoU loss function

We replace the GIoU loss function with YOLOX's original IOU regression loss function, keep other settings unchanged, and observe changes in model performance.

Table 3 shows the precision, recall, and mAP results of the ablation experiment.

Table 3. Ablation experimental result

	Precision/%	Recall/%	mAP/%
Baseline model	83.78	79.78	81.77
No swin-transformer	84.09	79.49	82.70
No ASFF	83.98	79.69	81.80
No GIoU	84.37	80.09	82.22
SwiF-YOLO	84.54	81.37	83.17

As shown in Table 3, for SwiF-YOLO, the performance decreases the most after removing the ASFF module, indicating that ASFF contributes the most to the model, followed by swin-transformer, and finally GIOU. Different combinations also positively optimize SwiF-YOLO. Overall performance. Compared with YOLOX-M, using Swin-Transformer to replace the backbone network and adding ASFF feature fusion has the greatest improvement in model performance, while the combination of Swin-Transformer and GIoU regression loss function has the smallest improvement in performance. The combination of the three

improved features has the greatest improvement in detection accuracy and performance. The improvement effect is the best.

4.5. Performance Comparison between SwiF-YOLO and other Detection Models

To verify the effectiveness of the proposed model, we compared it with other typical nodule detection methods published in recent years. Table 4 shows the comparison results with Faster R-CNN, Mask R-CNN, SSD, YOLO-v5, YOLO-v7 and YOLOX-M.

Table 4. Comparison experiment

Model	Precision/%	Recall/%	mAP/%
Faster R-CNN	81.76	82.45	82.40
Mask R-CNN	74.54	78.30	73.65
SSD	74.68	77.00	75.20
YOLO-v5	74.20	74.78	80.22
YOLO-v7	82.68	74.72	81.28
YOIOX-M	83.78	79.78	81.77
SwiF-YOLO	84.54	81.37	83.17

According to the experimental results, the following conclusions can be drawn. Compared with the two-stage Faster R-CNN, Mask R-CNN, and the one-stage SSD, YOLO-v5, YOLO-v7, and YOLOX-M models, although the recall rate of the SwiF-YOLO model is 1.08% lower than Faster R-CNN, its accuracy and average precision are better than Faster R-CNN, and significantly better than Mask R-CNN, YOLO-v5 and SSD models. In addition, compared with the benchmark model YOLOX-M, the SwiF-YOLO model improves accuracy, recall and mAP by 0.76%, 1.59% and 1.4% respectively. Although the recall rate of the SwiF-YOLO model is not outstanding, it has high accuracy and mAP.

5. Conclusion

This paper proposes a new method based on the YOLOx model, called SwiF-YOLO, to improve the accuracy and efficiency of pulmonary nodule detection. Swin Transformer was introduced in the basic model to replace the original

backbone network, and ASFF was added to the feature fusion layer to improve the model's perception of targets at different scales. Finally, the regression loss function IoU is replaced by GIoU. In the experimental results section, the experimental environment and the data sets used are first introduced, and then the experimental process is described in detail. Then, ablation experiments were conducted on the LUNA16 dataset to verify the effectiveness of the proposed detection method. In the ablation experiment, each improvement point of the model was removed one by one, including the Swin Transformer backbone network, ASFF feature fusion and GIoU loss function, and the performance changes of the model were compared. Finally, the model was trained on the public data set LUNA16, and compared with other mainstream target detection models in terms of detection accuracy, recall rate and mAP.

References

- [1] Sung H, Ferlay J, Siegel R L, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries[J]. CA: A Cancer Journal for Clinicians, 2021, 71(3): 209-249.
- [2] Liu K. Stbi-yolo: A real-time object detection method for lung nodule recognition[J]. IEEE Access, 2022, 10: 75385-75394.
- [3] Lu S, Yu Y, Yang Y. Retrospect and Prospect for Lung Cancer in China: Clinical Advances of Immune Checkpoint Inhibitors[J]. Oncologist, 2019, 24: S21-S30.
- [4] Henschke C I, Mccauley D I, Yankelevitz D F, et al. Early Lung Cancer Action Project: overall design and findings from baseline screening[J]. Lancet (London, England), 1999, 354(9173): 99-105.
- [5] Zhang G, Jiang S, Yang Z, et al. Automatic nodule detection for lung cancer in CT images: A review[J]. Computers in Biology and Medicine, 2018, 103: 287-300.
- [6] Haider W, Sharif M, Raza M. Achieving accuracy in early stage tumor identification systems based on image segmentation and 3D structure analysis[J]. Computer Engineering and Intelligent Systems, 2011, 2(6): 96-102.
- [7] Abubaker A, Qahwaji R S, Aql M, et al. Average row thresholding method for mammogram segmentation[J]. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference, 2005, 2005: 3288-91.
- [8] Girshick R. Fast r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2015: 1440-1448.
- [9] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
- [10] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision, 2017: 2961-2969.
- [11] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2019: 6569-6578.
- [12] Law H, Deng J. Cornernet: Detecting objects as paired keypoints [C]. Proceedings of the European conference on computer vision (ECCV), 2018: 734-750.
- [13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
- [15] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF international conference on computer vision, 2021: 10012-10022.
- [16] Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516, 2019.
- [17] Setio A a A, Traverso A, De Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge[J]. Medical image analysis, 2017, 42: 1-13.