A Tuberculosis Detection Method Using Attention and Sparse R-CNN

Xuebin Xu[1,2,3], Jiada Zhang[1,2,3*], Xiaorui Cheng[1,2,3], Longbin Lu[1,2,3], Yuqing Zhao[1,2,3], Zongyu Xu[1,2,3] and Zhuangzhuang Gu[1,2,3]

[1] School of Computer Science and Technology, Xi'an University of Posts and Telecommunications
Xi'an, 710121, China
[e-mail: xuxuebin@xupt.edu.cn, jiadaz1996@163.com, 617736425@qq.com, lulongbin@xupt.edu.cn, 2282788093@qq.com, 861953467@qq.com, 2499446557@qq.com]
[2] Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing
Xi'an, 710121, China
[3] Xi 'an Key Laboratory of Big Data and Intelligent Computing
Xi'an, 710121, China
[*]Corresponding author: Jiada Zhang

Received December 9, 2021; revised May 7, 2022; accepted May 28, 2022; published July 31, 2022

Abstract

To achieve accurate detection of tuberculosis (TB) areas in chest radiographs, we design a chest X-ray TB area detection algorithm. The algorithm consists of two stages: the chest X-ray TB classification network (CXTCNet) and the chest X-ray TB area detection network (CXTDNet). CXTCNet is used to judge the presence or absence of TB areas in chest X-ray images, thereby excluding the influence of other lung diseases on the detection of TB areas. It can reduce false positives in the detection network and improve the accuracy of detection results. In CXTCNet, we propose a channel attention mechanism (CAM) module and combine it with DenseNet. This module enables the network to learn more spatial and channel features information about chest X-ray images, thereby improving network performance. CXTDNet is a design based on a sparse object detection algorithm (Sparse R-CNN). A group of fixed learnable proposal boxes and learnable proposal features are using for classification and location. The predictions of the algorithm are output directly without non-maximal suppression post-processing. Furthermore, we use CLAHE to reduce image noise and improve image quality for data preprocessing. Experiments on dataset TBX11K show that the accuracy of the proposed CXTCNet is up to 99.10%, which is better than most current TB classification algorithms. Finally, our proposed chest X-ray TB detection algorithm could achieve AP of 45.35% and AP50 of 74.20%. We also establish a chest X-ray TB dataset with 304 sheets. And experiments on this dataset showed that the accuracy of the diagnosis was comparable to that of radiologists. We hope that our proposed algorithm and established dataset will advance the field of TB detection.

Keywords: Tuberculosis, Chest X-ray, Computer-aided diagnosis, Object detection, Attention.

1. Introduction

Tuberculosis (TB) is a global infectious disease with a high fatality rate. The World Health Organization estimates that there will be approximately 9.96 million new cases of tuberculosis and 1.21 million deaths globally in 2019 [1]. TB is highly contagious, spreads easily, and spreads rapidly from person to person. To effectively treat TB patients and reduce the risk of disease transmission, a rapid and accurate diagnosis of TB is very important. If patients are diagnosed with TB early and treated, their chances of survival can be significantly improved [2].

The sputum smear test is the gold standard for the diagnosis of TB. It tests the sputum to find Mycobacterium tuberculosis and diagnose TB. However, the long wait times for the process, and the inability of many hospitals and resource-constrained communities in many developing countries to afford such conditions, make this diagnostic method very limited [3-4]. Since TB has visual symptoms such as fibrosis, infiltration, mass, nodule, etc., professional radiologists can identify most symptoms on chest X-rays because chest X-ray imaging is cheaper and easier to obtain chest images. Therefore, chest X-ray is currently the primary method for diagnosing TB, and it is also the first choice for initial screening of lung diseases in most countries [5-9]. However, in actual work, some other lung diseases, such as pulmonary nodules, aseptic pneumonia, myocarditis, etc., have abnormalities similar to tuberculosis on chest X-rays (such as vague irregular lesions.) [10]. Therefore, if multiple symptoms appear together, it will bring great difficulties to the doctor in the diagnosis process. Moreover, when doctors examine chest radiographs, subjective differences, image quality, and fatigue caused by heavy work can significantly affect the diagnosis. Therefore, diagnosing and treating pulmonary tuberculosis is a time-consuming and challenging task [11].

In recent years, many researchers have been working to develop a computer-aided detection (CAD) system, hoping to use medical imaging and CAD systems for the initial diagnosis of TB. However, due to the complexity of chest radiograph data and the lack of pulmonary tuberculosis detection data sets, the existing CAD systems have low sensitivity and specificity in diagnosing pulmonary tuberculosis.

Therefore, this paper proposes an algorithm to identify and detect TB regions on chest X-ray images. The algorithm consists of two main network models: the Chest X-ray Tuberculosis Classification Network (CXTCNet) and the Chest X-ray Tuberculosis Detection Network (CXTDNet). We propose a channel attention mechanism (CAM) in CXTCNet and combine it with DenseNet. CXTCNet was used to reduce the false-positive probability of TB detection. Before TB detection, CXTCNet classifies chest X-ray images as healthy, TB, and unhealthy but non-TB (Sick). Then input the identified TB data into the detection network. CXTDNet is designed based on a sparse object detection algorithm (sparse R-CNN) to detect TB regions on TB data obtained from CXTCNet [12]. Moreover, we use CLAHE to pre-process the data. Finally, it is validated on TBX11K and our established dataset, proving that the algorithm in this paper can accurately locate the TB area on chest radiographs.

The rest of the paper is structured as follows: Section 2 describes related work on tuberculosis detection. Section 3 describes the TBX11K dataset, and our established dataset TBX304 Section 4 describes the preprocessing method and the methodology of this study, etc. The experimental results and comparative analysis are presented in Section 5. Finally, the paper is summarized and concluded with future directions in section 6.

2. Related Work

Before the rise of deep learning, the classification of TB and non-TB cases in chest radiographs was dominated by traditional methods. These methods first extract features by hand and then combine them with supervised learning algorithms for identification. Jaeger et al. [13] used a graph segmentation method to segment lung regions and obtained a set of image features from these regions. Then, they used a support vector machine (SVM) to classify the chest radiographs into TB and non-TB. Jeyavathana et al. [14] compared three ROI feature extraction methods: Local Binary Pattern (LBP), Histogram of Gradients (HOG), and Tamura Texture Features. They proposed a method to extract LHTGF features (LBP, HOG, LHT, and Gabour filters) from local windows to identify tuberculosis in the chest radiograph. Since traditional methods are based on manual extraction of features to process medical images, it requires algorithm designers to have a wealth of medical knowledge to extract high-quality features with sufficient discrimination. In addition, the performance of traditional machine learning classifiers is limited by the quality and distribution of training samples and is prone to overfitting due to insufficient training samples, resulting in poor recognition performance of the algorithm. With the rapid development of a convolutional neural network (CNN), CNN has gradually been used to identify tuberculosis in chest X-ray images [15-20]. For example, Hwang et al. [21] designed a deep-learning-based automatic detection (DLAD) algorithm. Experiments used a dataset with 54,221 normal chest radiograph images and 6768 chest radiograph images of active tuberculosis for validation. The final results showed that the DLAD algorithm performed well in detecting active tuberculosis, outperforming most physicians, including thoracic radiologists. Rahman et al. [22] used image preprocessing, image segmentation and classification techniques to detect TB and evaluated the performance of 9 different CNNs in identifying TB. The simulations showed that DenseNet201 achieves the best results in chest X-ray images after U-Net segmentation, with accuracy, precision, and recall rates of 98.6%, 98.57%, and 98.56%, respectively. Ayaz et al. [23] proposed a tuberculosis detection technique that combines handcrafted features with deep CNN features via ensemble learning. They evaluated the proposed method using the Montgomery and Shenzhen datasets. The simulations show that the maximum accuracy of the Montgomery dataset is 93.47%, and the maximum accuracy of the Shenzhen dataset is 90.6%.

However, these studies only classified chest radiographs and did not locate TB areas. Currently, there are few research works related to the detection of TB area. The main reason is that tuberculosis data is too sensitive and violates patients' privacy. Therefore, few publicly available TB detection datasets bring significant obstacles to this research [24-25]. To solve this problem, Yun Liu et al. [26] established a TB detection dataset: TBX11K. This dataset contains 11200 X-ray images. In addition, the corresponding locations with TB areas in the TB samples are annotated. Finally, these improved detection algorithms are simulated on TBX11K and used as a baseline for future research. The proposed TBX11K dataset and reference baselines are expected to advance research in CAD systems and design better CAD systems through new powerful deep networks.

3. Datasets

We used two datasets, a public tuberculosis X-ray (TBX11K) dataset and one diagnosed, labeled, and created by several radiologists from the Shaanxi Provincial Tuberculosis Control Hospital. The public dataset was mainly used for experimental comparison with other existing studies. The dataset we established was used for comparison with radiologists, making this

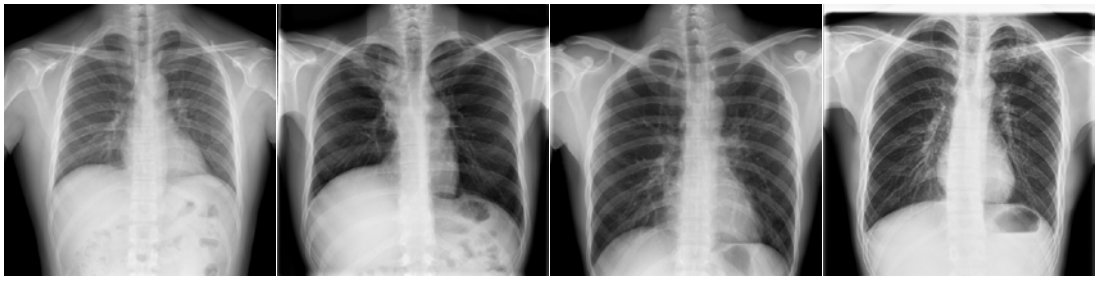
study more convincing and reliable. **Table 1** below shows the statistics of the number of datasets used in this paper, and the different datasets will be described in detail below.

Table 1. Number of datasets used in this paper

Dataset	Category	Number (sheets)	Total (sheets)
TBX11K	Health	5000	11494
	diseased but non-TB (Sick)	5000	
	TB	1190	
we established	TB	304	

3.1 TBX11K

TBX11K is a TB chest X-ray dataset established by Liu et al. It consists of 11,200 X-rays, of which 5,000 are healthy samples, 5,000 are unhealthy but non-TB samples, and 1,200 are TB samples. Also, the TB samples contained different types of TB. Among them, there are 924 active cases, 212 latent cases, and 54 cases with both active and latent cases. The type of the remaining 10 cases is still to be determined. The pending of all images was 512 × 512. For images with TB manifestations, in addition to using boxes to locate TB areas, the type of each TB area is also differentiated. The 10 indeterminate cases out of the 1200 cases with TB manifestations that could not be recognized as TB type under today's medical conditions were not used in this paper. In conclusion, TBX11K can be divided into the following categories, healthy, sick, and TB, of which TB is divided into latent and active. In **Fig. 1**, we show some of the data from the dataset TBX11K.



(**a**) Health (**b**) Sick (**c**) Active TB (**d**) Latent TB

Fig. 1. Shown is a partial X-ray image of the dataset TBX11K. a, b, c and d are the image data of Health, Sick, Active TB and Latent TB types respectively.

3.2 Established dataset

In order to validate proposed detection method in a real environment, a new TB dataset was created--TBX304. This dataset was established in collaboration with the Shaanxi Provincial Tuberculosis Hospital. Several radiologists diagnosed and labeled it with many years of clinical experience, which has a high degree of authenticity and accuracy. The dataset has 304 chest X-ray images, all of which are active tuberculosis and marked with the corresponding bounding boxes of tuberculosis areas. The initial resolution of all X-ray images was around 3000×3000. However, to improve the processing speed and ensure that the images do not lose features due to excessive compression, the resolution of all images was adjusted from 3000×3000 to 512×512. Some of the data from our newly created dataset are shown in **Fig. 2**.

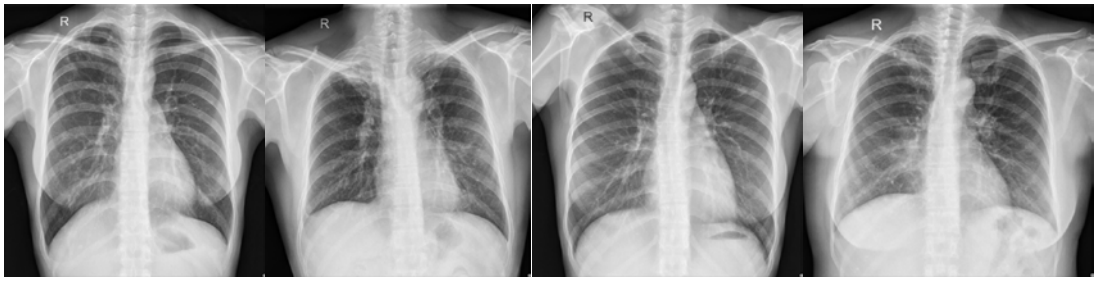


Fig. 2. Shown is a partial x-ray image of TBX304.

4. Methods

4.1 Algorithm structure

Since there are many chest X-ray images of diseased but non-TB (Sick), the direct use of TB region detection algorithms may lead to misdiagnosis [10]. To solve this problem, we design a high-precision classification algorithm to classify chest X-ray images into healthy, TB, and sick. Efficient screening of TB area detection was carried out first, which helped reduce false positives in detection results and improve detection performance. TB is divided into latent TB and active TB. Latent TB means that the patient had TB before, but it has been cured and has no health consequences or contagiousness. On the other hand, active TB means that the patient has TB and is contagious, so the distinction between the two is important to the doctor's diagnosis [27]. Therefore, this paper not only detects TB areas but also identifies TB types, to assist physicians in more accurate diagnosis and treatment. **Fig. 3** shows the detection flowchart of the chest X-ray TB area detection algorithm proposed in this paper. The overall process is as follows: first, the dataset is preprocessed with data, and then the chest X-ray images are classified into three categories by CXTCNet, i.e., healthy, sick, and TB. When the recognition result is healthy or sick, the result is output directly; when the recognition result is TB, the chest X-ray image is input to CXTDNet to detect the TB.

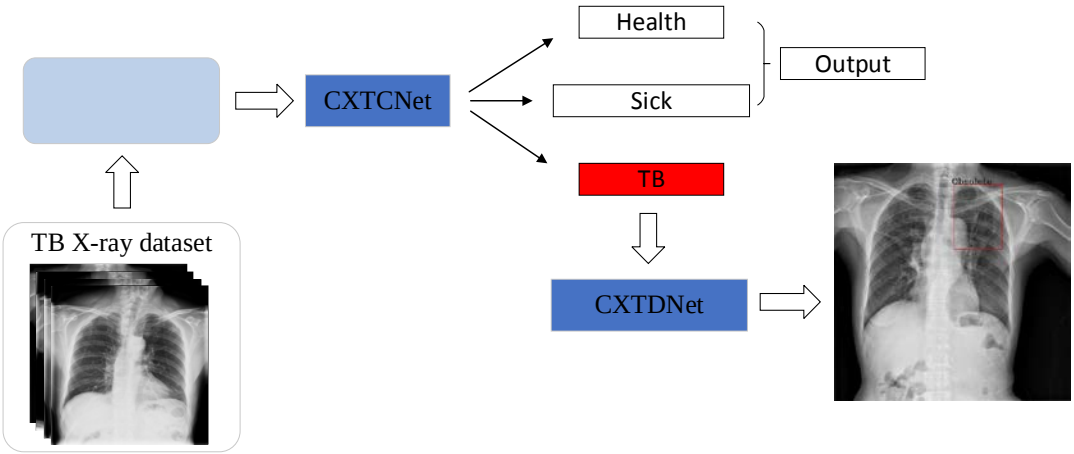


Fig. 3. The architecture of the TB detection algorithm proposed in this paper. CXTCNet is Chest X-ray TB Classification Network. CXTDNet is Chest X-ray TB area Detection Network

4.2 Data preprocessing

To improve the image quality, we used Contrast Limited Adaptive Histogram Equalization (CLAHE) to reduce the noise level and enhance the contrast of the medical image [28]. After preprocessing the image by CLAHE, it is more helpful to extract some essential features when executing the algorithm steps later and analyze the presence of tuberculosis lesions in that image by these features. CLAHE is a method to improve the low contrast problem of digital images. It has been shown that CLAHE is well suited for biomedical images such as mammograms, where it can improve image quality by removing noise [29-31]. Therefore, in this paper, CLAHE is introduced into the preprocessing stage of data to preprocess the data and improve the image's contrast. CLAHE limits the magnitude of contrast enhancement by limiting the height of the local histogram to avoid amplification of noise and excessive contrast enhancement. The algorithm redistributes the histogram part that exceeds the trimming limit to other parts of the histogram by setting a trimming limit value. Furthermore, we can limit the slope of the transform function during local histogram equalization to avoid the problem of noise amplification caused by the over-enhancement of narrowband pixels.

CLAHE solves some of the problems caused by standard Histogram Equalizatio (HE): (1) areas with too much contrast enhancement become noisy (2) some areas become darker or brighter after adjustment, resulting in more details lost. Moreover, CLAHE also solves the problems such as image distortion brought by Adaptive histogram equlization (AHE). CLAHE preprocesses the data used in this paper at the beginning, and the preprocessed images are shown in **Fig. 4**. The image on the left is the original image, and the one on the right is the image obtained after CLAHE preprocessing. As shown in **Fig. 4**, the white area in the box in the left image is the tuberculosis lesion. The contrast between this location and the lung shadow in the original image is low, and the characteristics of the tuberculosis lesion are not very obvious. In contrast, the image after CLAHE preprocessing can see the image details with high contrast and at the same time has good noise suppression. It is known that preprocessing with CLAHE can enhance the image details while avoiding noise amplification and has a good enhancement effect.

Fig. 4. Comparison of the original image (left) and the image after CLAHE preprocessing (right).

4.3 Chest X-ray TB Classification Network (CXTCNet)

In recent years, deep learning has evolved rapidly, and with it many convolutional neural network (CNN) models, including LeNet5 proposed in 1998, and later AlexNet, DenseNet, SENet, etc [32-36]. Studies have proved that CNN has achieved great success and has been widely adopted by the computer vision community, suitable for tasks such as image classification and image detection [37-39]. A CNN is structured mainly with several convolutional (Conv) and pooling layers, and at least one fully connected (FC) layer is connected at the end. Among them, the convolution layer has multiple convolution kernels with trainable weights. Multiple feature maps are finally generated by convolving the image with each convolution kernel and adding a bias to the convolution layer. The pooling layer is a non-linear down-sampling process that can save the relevant information about the task, increase the receptive field of the feature map, and remove irrelevant details. The previous convolutional and pooling layers are for feature extraction of the image, and the fully connected layer is to classify the extracted features. The classification is realized by mapping the features to neurons.

By analyzing the recognition effects of existing classification methods, it is found that these methods perform well in many image classification scenarios. However, for complex application scenarios such as medical images, the recognition accuracy that can be achieved is not very good, especially for image data of chest radiographs where multiple diseases exist, and the lesions of different diseases are relatively similar. Therefore, we did not choose to use the existing CNN model directly for chest radiograph classification but by analyzing the existing TB classification studies and classification networks. Then DenseNet was chosen as the backbone, a channel attention module (CAM) was proposed, which led to the design of a chest radiograph tuberculosis classification network (CXTCNet). The image can be convolutional operated to get the feature map. To obtain the channel feature on the feature map so that the neural network can learn the relationship between the feature map channels through backpropagation, we designed CAM. CAM is composed of different pooling operations, fully connected layers, and activation functions, as shown in **Fig. 5**.

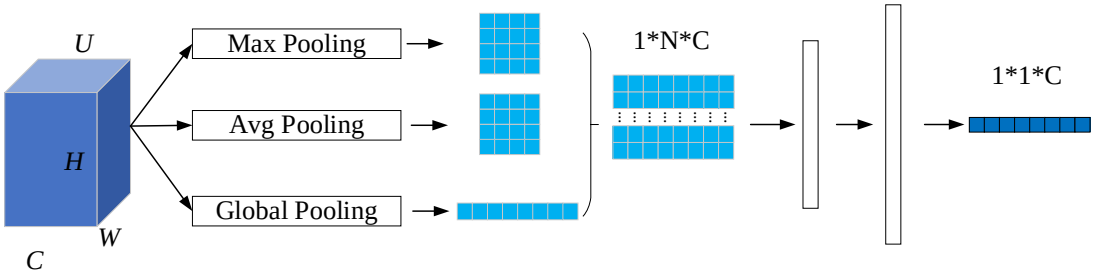


Fig. 5. Structure of the Channel Attention Module (CAM).

When the feature map enters CAM, CAM first uses three different pooling operations to extract features of the feature map. Then use the first FC to map the feature information and at the same time, play the role of dimensionality reduction. The activation function of this FC uses ReLU. Finally, the second FC is used to upgrade the dimension, and the activation function Sigmiod is used to limit the output to (0, 1) as the weight of each channel of the feature map. CAM uses the learned weights to limit or play the importance of the corresponding channels and control the influence of different channels on the final output result. In order to obtain the weight coefficients of each channel, the global average pooling

(GAP) is used to extract the global feature information of each channel of the feature map, and the value calculated by GAP is the global distribution of the corresponding feature channels. As shown in **Fig. 6**, GAP transforms the feature map of size W*H*C into a weight coefficient of 1*1*C. The calculation formula of GAP is:

$$F_{GAP} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} u_C(i, j) \tag{1}$$

F_{GAP} is the result of the GAP, W and H are the width and height of the feature map, respectively. C is the channel number of the feature map, and $u_C(i, j)$ is the value of the i-th row and j-th column in the c-th channel of the input feature.

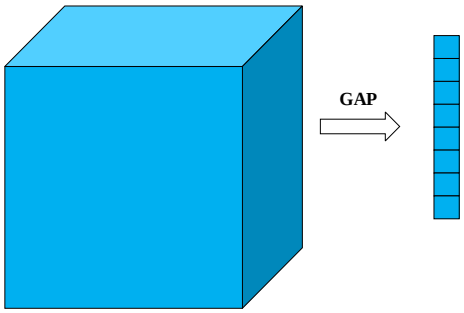


Fig. 6. Extract the information of feature maps using GAP.

However, only the learning of global features will result in the loss of some information, so we have added local average pooling and local maximum pooling to retain more image feature information to improve the final recognition performance of the network. Moreover, to reduce the number of parameters of CNN in the calculation process, we limit the results of local pooling. When the size of the feature map is greater than or equal to 128*128, the pooling window is 7*7 and the stride is 7; when the size of the feature map is greater than or equal to 64*64, the pooling window is 5*5 and the stride is 5; When the size of the graph is smaller than 64*64, the pooling window is 3*3, and the stride is 3. After the above three pooling operations, a feature map of size 1*N*C can be obtained, and then the feature matrix is input to the first FC and the activation function ReLU. In this way, the complex correlation between the channels can be well fitted. The selection of fewer neurons is also beneficial in reducing the number of parameters and the amount of calculation. The resulting feature matrix with a size of 1*N*C will be input to the second fully connected layer and the activation function sigmoid, and the number of neurons is the same as the number of channels. Therefore, a feature matrix with a size of 1*1*C can be obtained, and each value is between (0, 1), representing the weight ratio of each channel. Finally, a multiplication operation is used to multiply the obtained weight coefficients with the corresponding feature channels in the feature map. The weight coefficients are used to suppress or play the importance of the corresponding channels. Furthermore, through backpropagation, the weight coefficients corresponding to unimportant channels are reduced, and the weight coefficients corresponding to more important channels are increased, so that the network can learn more features.

Dense Convolutional Neural Network (DenseNet) eliminates the problems of gradient disappearance and gradient explosion by establishing dense connections between different layers. At the same time, it uses the shallow features of the network model to enable the network to learn more feature information and realize Feature reuse. Compared with other deep convolutional neural network models such as ResNet, DenseNet reduces a large number of training parameters, effectively suppressing the overfitting phenomenon in the model

training process, making DenseNet stronger in generalization ability and better performance. Therefore, we use DenseNet as the backbone and cooperate with CAM to fuse the image's spatial feature information and channel feature information. In order to ensure that sufficient feature information is learned, we place the CAM behind the Dense Block to avoid the loss of feature information caused by operations at the transition layer. The specific connection is shown in **Fig. 7**. After the feature map passes through the Dense Block, a new feature map is obtained. The calculation formula of the Dense Block is:

$$H = F(\left[h_0,\ h_1,\cdots,\ h_{l-1}\right]) \tag{2}$$

Where H is the output of Dense Block and $[h_0, h_1,\ldots,h_{l-1}]$ is the feature map of the output of layers 0 to l-1, respectively. Then, the feature map is input to CAM to obtain the weight matrix, which is calculated by：

$$X = F_{CAM}(H) = s(W_2,(r(W_1,(F_{GAP}(H),F_{MAX}(H),F_{AVG}(H))) \tag{3}$$

Where X is the weight coefficient of CAM output, F_{GAP}, F_{MAX}, F_{AVG} are the results of feature map obtained by global average pooling, local maximum pooling and local average pooling, respectively. W_1 is the neuron weight of the first FC, r is the activation function ReLU, W_2 is the neuron weight of the second FC, and s is the activation function Sigmoid. Finally, the weight matrix is multiplied by the input feature map, and the calculation formula is as follows:

$$\tilde{H} = F(H,\ X) \tag{4}$$

Where $\tilde{H}$ is the output result and F represent the corresponding channel multiplication of the feature map output by the Dense Block and the weight coefficient obtained by the CAM.

We added CAM to each Dense Block, because the module has a small amount of parameters, so it can rudece the calculation burden of the network. Finally, the Softmax classifier is used for classification, and the formula of the loss function is as follows:

$$L_{cla} = -\sum_{i}^{k=3} y_i \log a_i = -\sum_{i} y_i \log(\frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}}) \tag{5}$$

Where y is the true value, a is the value obtained by Softmax, and k is the number of categories.

The classification network recognizes the chest radiographs of patients suffering from tuberculosis in advance, thereby preventing other lung diseases with similar abnormalities to tuberculosis from affecting the test results, reducing false positives in tuberculosis area detection, and improving the detection performance.

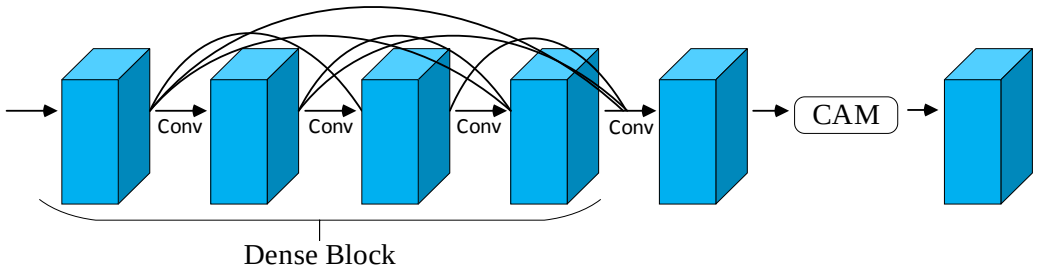


Fig. 7. CAM and DenseNet connection structure diagram.

4.4 Chest X-ray TB area Detection Network (CXTDNet)

Object detection is an important branch of computer vision. With the deepening of neural network theoretical research and the substantial increase in hardware GPU computing power,

it has become a hot spot in global artificial intelligence research [40-41]. Object detection was mainly set up with dense candidate boxes. For one-stage detectors, such as YOLO, SSD, RetinaNet. [42-47], these detectors would directly predict the classification and position of thousands of dense anchor boxes placed in the image space, and their candidate box settings resulted in the presence of a large number of hyperparameters. These hyperparameters include the number of anchors at each position, the size of the anchor, and the aspect ratio of the anchor. For two-stage detectors, such as the Faster RCNN, the foreground and background regions are first classified from a predetermined dense set of anchor boxes, resulting in sparse region proposals, which are then put into the later network for finer classification and position regression [48]. However, dense candidate boxes are set in the model in the one-stage and two-stage, which is a significant burden to the detection network. Furthermore, these dense object detection algorithms produce many similar results, so at the end a post-processing with non-maximum suppression (NMS) is required [49]. Recently, the DETR network, which migrates transformers to computer vision tasks, has received widespread attention. This network transforms the target detection task into a task of set prediction, using a transformer coder-decoder structure and a bilateral matching approach to obtain prediction results directly from the input image [50]. Unlike other detection methods, DETR has no proposal, no anchor, no center, and no cumbersome NMS. It directly predicts the detection boxes and categories. It uses the Hungarian algorithm of bipartite graph matching to achieve the task of object detection with a clever combination of CNN and transformer. Nevertheless, DETR has many disadvantages, such as slow convergence and high inference memory usage. Therefore Peize Sun et al. proposed sparse R-CNN, Sparse R-CNN is extremely simple, with no need to set an annoying dense anchor, no RPN, no complex post-processing and NMS, no need to balance the RPN and fast RCNN training process carefully, and no hard-to-tune hyperparameters. The effect is better than Faster R-CNN and the convergence speed is much faster than DETR, so we choose Sparse R-CNN to be the detection network. The sparse R-CNN is a simple, unified network composed of a backbone network, a dynamic instance interaction head, and two task-specific prediction layers, as shown in **Fig. 8**. In the initial input of Sparse R-CNN, in addition to the input image data, a set of proposal boxes and corresponding proposal features are also input. This set of proposal boxes and proposal features is optimized along with other parameters in the network.

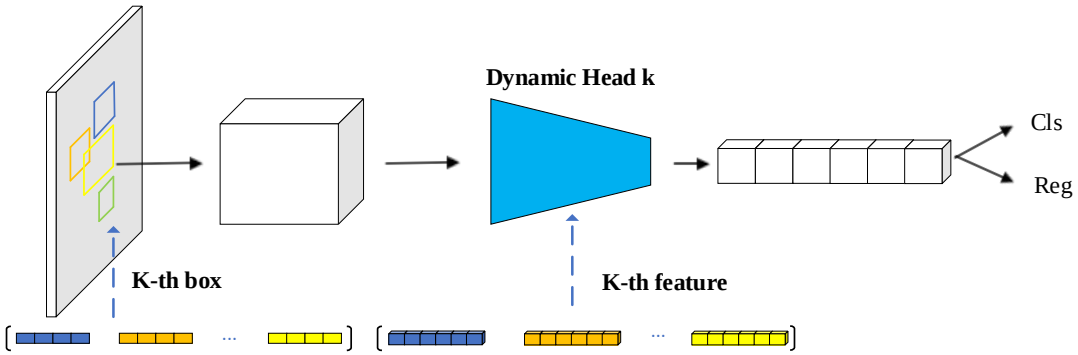


Fig. 8. Structure of Sparse R-CNN.

The backbone network used by Sparse R-CNN is a FPN based on the ResNet architecture [51]. After analyzing the appeal classification experiments, it is known that our proposed CXTCNet can extract enough image feature information, so we change the FPN based on ResNet architecture to CXTCNet-based FPN as our backbone network.

The proposal boxes of sparse R-CNN are not from FPN, but are used as region proposals through a set of learnable proposal boxes (n4) [52]. The value range of the proposal box is (0, 1), and the number of the proposal box is determined by the hyperparameter d, which represents the d group of proposal boxes. Each group of proposal boxes has 4 values, representing the center coordinate, height, and width, respectively. These learned proposal boxes can be viewed as initial guesses for the regions most likely to contain objects in the image. A back-propagation algorithm will update the parameters of proposal boxes during the training process.

Although the proposal box is a concise and clear way to describe an object, it can only represent the rough positioning of the object. It cannot represent the more detailed feature information in the image, such as the posture and shape of the object. Therefore, the author added a proposal feature, whose size is N*d, to enrich the latent information of the feature. The number of proposal features is the same as the number of proposal boxes, and there are d groups of proposal features when there are d groups of proposal boxes. There are N values in each set of proposed features, and the default value of N is 256.

Fig. 9 shows the architecture of the dynamic instance interactive header. Sparse R-CNN first uses the RoIAlign operation to extract the features of each proposal box. Then each RoI feature is fed into its exclusive head for object location and classification, where each head is conditioned on a specific proposal feature. Finally, each RoI feature will interact with the corresponding proposal feature to filter out ineffective bins and outputs the final object feature. For lightweight design, continuous 1×1 convolution with ReLU activation function is used to realize the interactive process. To obtain a more discriminative feature, each feature is convolved with the RoI area feature to obtain a more discriminative feature. The final regression prediction uses FC and ReLU activation functions.

$$Loss = \lambda_{cls} \bullet L_{cls} + \lambda_{L1} \bullet L_{L1} + \lambda_{giou} \bullet L_{giou} \tag{6}$$

Where L_{cls} represents the classification loss, and the focal loss is used here. The regression loss for objects is the weighted sum of the L1 loss and the generalized iou loss, i.e. L_1 and L_{giou} in Eq. (6). $\lambda_{cls,}$, λ_{L1}, and λ_{giou} denote the coefficients of L_{cls}, L_1, and L_{giou}, respectively [53].

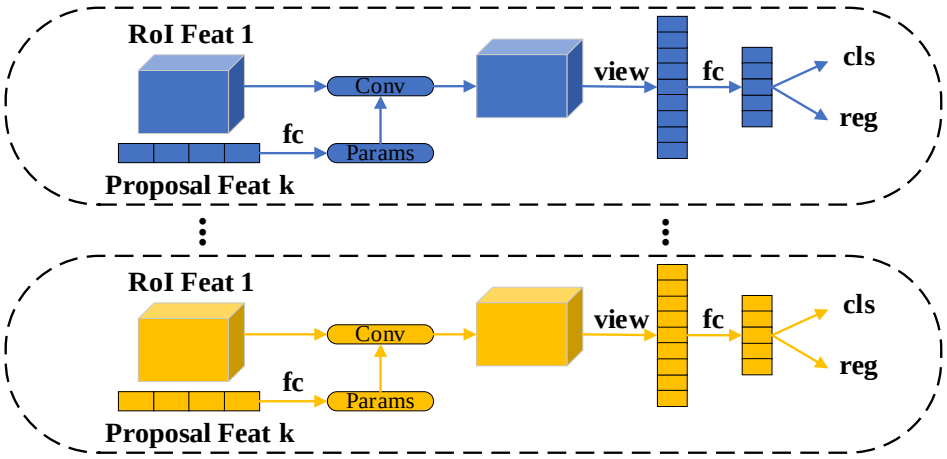


Fig. 9. The structure of the Dynamic instance interactive head.

5. Experimental results and analysis

5.1 Experimental analysis of CXTCNet

5.1.1 Evaluation metrics

Since the classification network is used to identify healthy, sick and TB, the evaluation metrics used in this experiment are Accuracy, Precision, Recall and F1 score.

5.1.2 Parameters adjustment

Due to the unbalanced distribution of the data in the three categories of TBX11K, there are 5000 images for both health and sick, but only 1200 for TB, which will lead to overfitting if training is performed directly. Furthermore, we only selected 1190 radiographs with TB manifestations out of 1200, and the uncertain 10 we screened out. Therefore, we randomly took 1190 images from each class of data and performed data enhancement to have more balanced data. Then the enhanced data is randomly divided into the training set, validation set and test set, with a ratio of 6:2:2. 8 V100 GPUs were used for experiments, and the deep learning framework was PyTorch. Through experimental verification and analysis, the optimal parameter settings are obtained. **Table 2** shows the experimental results using different parameters. The optimizer used is Adam. From **Table 2**, when the learning rate is 0.0001, the input size is 512*512, the batch size is 16, and the epoch is 500, the best classification performance is achieved with 99.10% accuracy, 99.00% precision, and 99.25% recall.

Table 2. Experimental results of different parameters

Learning rate	Input size	Batch size	Epoch	Accuracy	Precision	Recall	F1
0.001	512*512	16	500	96.07	96.53	96.92	96.73
0.0001	512*512	16	500	**99.10**	**99.00**	**99.25**	**99.12**
0.0001	512*512	16	300	98.20	97.85	98.27	98.06
0.001	224*224	16	500	95.97	95.70	96.05	95.87
0.0001	224*224	16	500	97.88	97.32	98.13	97.72
0.0001	512*512	8	500	98.13	97.72	98.54	98.13
0.0001	512*512	16	1000	98.87	98.54	98.93	98.73

5.1.3 Validation of CAM

To verify the effect of CAM on the algorithm's performance, we experimentally compared the algorithm with and without CAM. First, we experimentally verified the algorithm's performance without CAM and adjusted the model parameters to obtain the optimal parameter settings: optimizer is Adam, the learning rate is 0.0001, the input size is 512*512, the batch size is 16, and the epoch is 500. **Fig. 10** shows the accuracy change curves of the models with and without CAM in the training process. As shown in **Fig. 10**, when the classification network does not incorporate CAM, the accuracy during training is about 93%, and the fluctuation of accuracy is high. The accuracy on the test set is also only 91.87%, and the precision and recall rates are also low. When the classification network is added to the CAM, the loss steadily decreases during the training process, and the accuracy rate gradually increases. Finally, the accuracy rate on the test set is also significantly improved, as high as 99.10%, and the precision and recall have reached 99.00% and 99.25% respectively. Therefore, the addition of CAM enables the model to learn not only the spatial features, but also the channel features. This module significantly improves the performance of the classification network and achieves the purpose of accurately identifying different types of chest X-ray images.

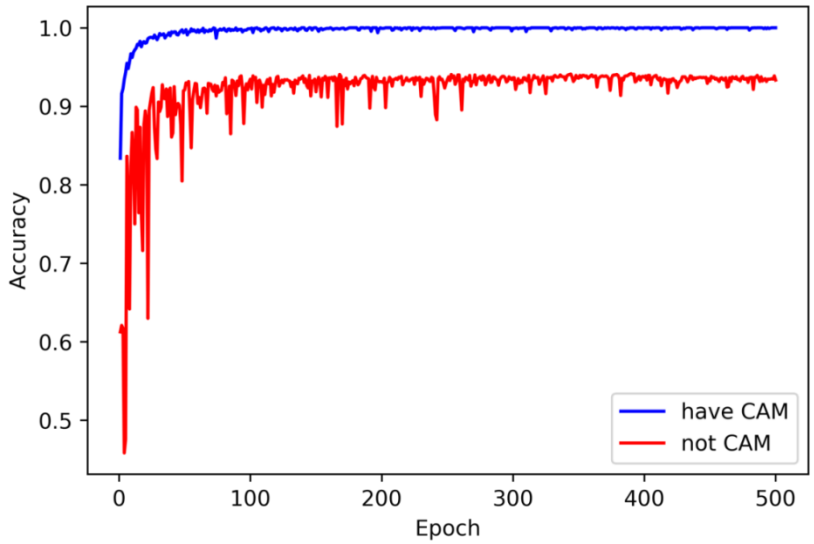


Fig. 10. Comparison of accuracy between adding CAM to the algorithm and not adding CAM.

5.1.4 Comparison with other classification algorithms

Our proposed algorithm (CXTCNet) has also been compared with other mainstream classification algorithms, using classification network models such as ResNet, DenseNet, and SENet. The following experimental results are finally obtained through constant adjustment of hyperparameters, as shown in **Table 3**. First, using resnet50 to experiment on the dataset and by adjusting the parameters, the highest accuracy of 87.91%, accuracy of 87.79% and recall of 87.05% were obtained, which is not as good as the performance of CXTCNet. The results of resnet101 are not excellent either. The highest accuracy rate is 90.15%, the accuracy rate is 91.35%, and the recall rate is 87.56%. DenseNet121 and DenseNet169 is slightly better, among which DenseNet169 is relatively good, with accuracy, precision and recall rates of 92.87%, 93.79% and 88.84% respectively. Finally, we also used SENet, an attention mechanism, for experimental validation. Experiments show that SENet outperforms the previous ResNet and DenseNet with an accuracy of 95.32%, precision of 96.23%, and recall of 94.85%. Although SENet is good, it is still 3.78 percentage points lower than CXTCNet. Therefore, we know that our proposed CXTCNet can accurately identify the features in different types of chest X-ray images and classify these features. Its performance is better than most existing classification network models.

Table 3. Comparison results of CXTCNet and other classification algorithms (%)

Method	accuracy	precision	recall	F1
ResNet50	87.91	87.79	87.05	86.92
ResNet101	90.15	91.35	87.56	89.41
DenseNet121	91.07	90.57	91.73	91.46
DenseNet169	92.87	93.79	88.84	91.25
SENet	95.32	96.23	94.85	95.54
CXTCNet (Our Method)	**99.10**	**99.00**	**99.25**	**99.12**

This paper also compares with some of the more advanced pulmonary tuberculosis identification algorithms. The following TB classification algorithm mainly uses the Shenzhen

and Montgomery County datasets. Therefore, when comparing with these algorithms, the datasets used in this paper are also these two datasets. Jaeger et al [12] used a variety of machine learning techniques to classify chest radiographs in these two datasets. Among them, the maximum accuracy rate of the Montgomery County dataset in the United States is 78.3%, and the accuracy rate of the Shenzhen dataset is 84%. Vajda et al. [54] proposed a fully automated chest X-ray system. The maximum accuracy rate obtained by the system on the Montgomery County dataset in the United States is 84.75%, and the maximum accuracy rate obtained on the Shenzhen dataset is 97.03%. Pasa et al. [55] conducted experiments on both datasets simultaneously and achieved a maximum accuracy of 92.5%. Tasci et al. [56] proposed a voting and preprocessing variant-based ensemble CNN model for TB detection. The accuracy of the proposed method on the Montgomery and Shenzhen datasets is 97.500% and 97.699%, respectively. Guo et al. [57] proposed an ensemble process to detect and localize tuberculosis using deep learning. The maximum accuracies obtained by this method on the Montgomery and Shenzhen datasets are 95.49% and 98.46%, respectively. The classification algorithm proposed in this paper obtained the maximum accuracy rates of 98.93% and 99.07% in the Shenzhen and Montgomery datasets, respectively. Furthermore, the accuracy rate obtained after merging the two datasets also reached 98.21%. The detailed comparison results are shown in Table 4. Compared with the work of Jaeger et al., the method proposed in this paper greatly improves the performance of TB classification on chest radiograph.

Table 4. Comparison of the accuracy of the classification algorithm proposed and other pulmonary tuberculosis identification algorithms on different datasets (%)

Method	Shenzhen dataset	Montgomery dataset	Shenzhen and Montgomery dataset
Jaeger et al. (2013) [12]	84	78.3	-
Ayaz et al. (2021) [22]	90.6	93.47	-
Vajda et al. (2018) [54]	97.03	84.75	-
Pasa et al. (2019) [55]	-	-	92.5
Tasci et al. (2021) [56]	97.699	97.5	-
Guo et al. (2020) [57]	98.46	95.49	
CXTCNet (Our Method)	**98.93**	**99.07**	**98.21**

5.2 Experimental analysis of CXTDNet

5.2.1 Evaluation metrics

For assessing of TB area detection, our evaluation metrics are consistent with those used by COCO, with the main values being AP and AP50. AP is the Iou threshold in the interval 0.5 - 0.95, and AP is calculated at 0.05 intervals and averaged thereafter. AP50 is the AP with an IoU threshold of 0.5 [58].

5.2.2 Parameters adjustment

In CXTDNet, we first trained using the default parameters of sparse R-CNN, i.e., ResNet50 was used as the backbone, the optimizer was AdamW and the weight decay was 0.0001, the batch size was 16, all models were trained on 8 GPUs, and the deep learning framework was PyTorch. The default number of iterations is 100 epochs, and the initial learning rate is set to 2.5*10-5, divided by 10 at epoch 80 and 90, respectively. The input size of the image is 512*512, and the default number of proposed boxes and proposed features are 100 and 100

respectively. Then the parameters were adjusted to find the optimal parameter settings for CXTDNet. The different results caused by different parameters are shown in **Table 5**. It is known from the experiments that the parameters that can have a significant impact on the results are batch size, epochs, learning rate, input size, and the number of proposed boxes. From **Table 5**, the optimal parameters of CXTDNet are set as follows: batch size is 16, epochs are 110, the learning rate is 2.5*10-5, the input size is 512*512, and the number of proposed boxes is 300. Furthermore, because no other two types of data were included, i.e., health and sick, CDTDNet achieved very good results. The final AP obtained was 45.35%, and the AP50 was 74.20%. **Fig. 11** shows the detection results of CXTDNet. Although the overall results were good, the detection of latent TB was poor, with an AP of only 30.78%. There are only 212 latent TB chest X-ray images in the dataset, while there were 924 active TB chest X-ray images, and the data set also includes two types of chest radiographs that co-exist. However, considering the high accuracy of detection for TB areas, it has been possible to reach the role of assisting doctors in diagnosis. So this data imbalance issue will continue to be looked at in later studies.

Table 5. Experimental results of different parameters of CXTDNet

Input size	Proposed boxes	Batch size	Learning rat	Epochs	AP(%)	AP50(%)
512*512	100	16	2.5*10-5	100	34.21	65.22
224*224	100	16	2.5*10-5	100	36.64	67.53
512*512	200	16	2.5*10-5	110	39.31	69.71
512*512	300	8	2.5*10-5	110	35.90	66.28
512*512	300	16	2.5*10-5	100	**45.35**	**74.20**
512*512	300	16	1.0*10-5	110	38.18	67.70
512*512	300	16	2.5*10-5	120	40.85	72.68

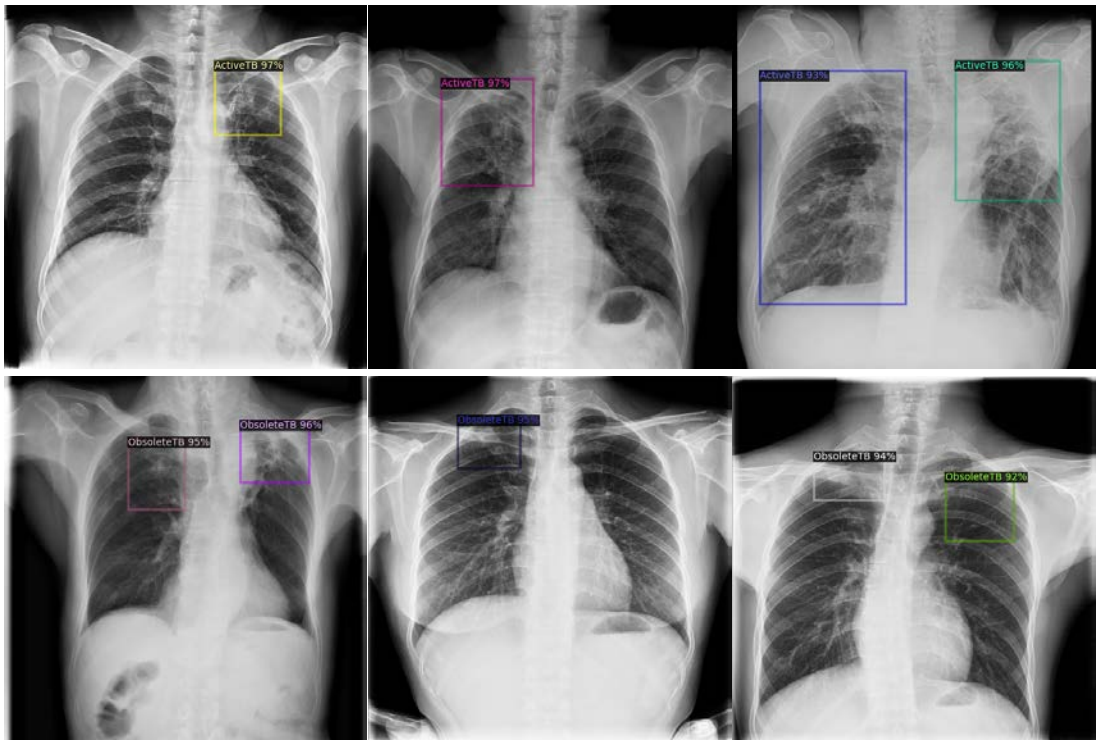


Fig. 11. The detection results of CXTDNet are shown here. The marked boxes in the figure show the TB areas detected by CXTDNet, with the corresponding TB types and the calculated confidence.

5.2.3 Validating the role of CXTCNet

In order to validate the enhancement brought by the classification network (CXTCNet) to the final detection results, we performed a set of validation experiments. Chest X-ray images were placed directly into Sparse R-CNN for training and testing, using the same backbone and parameter settings as described above. **Fig. 12** shows the results of this part of the experimental comparison. From **Table 6**, we can see that the detection results obtained by putting the chest X-ray images of healthy, sick and TB into Sparse R-CNN directly are not very satisfactory. AP is only 25.74%, and AP50 is only 63.15%. In contrast, our proposed algorithm for TB detection based on the attention mechanism and Sparse R-CNN achieves an AP of 44.90% and an AP50 of 73.28%. This analysis is mainly due to the similarity of abnormalities such as aseptic pneumonia, myocarditis, exudates, infiltrates, masses, nodules, and so on in chest radiographs with tuberculosis, which causes many misclassifications. Due to the presence of these pathologies, directly using the target detection algorithm would raise the false positive rate. With our proposed Chest X-ray Tuberculosis Classification Network (CXTCNet) for early screening, the Chest X-ray Tuberculosis Detection Network (CTXDNet) can avoid this problem and, thus, precisely local TB areas.

Table 6. Results of experimental validation using Sparse R-CNN on TBX11K and our proposed TB detection algorithm using attention mechanism and Sparse R-CNN on TBX11K

Method	AP(%)	AP50(%)
Sparse R-CNN	25.74	63.15
Our Method	**44.90**	**73.28**

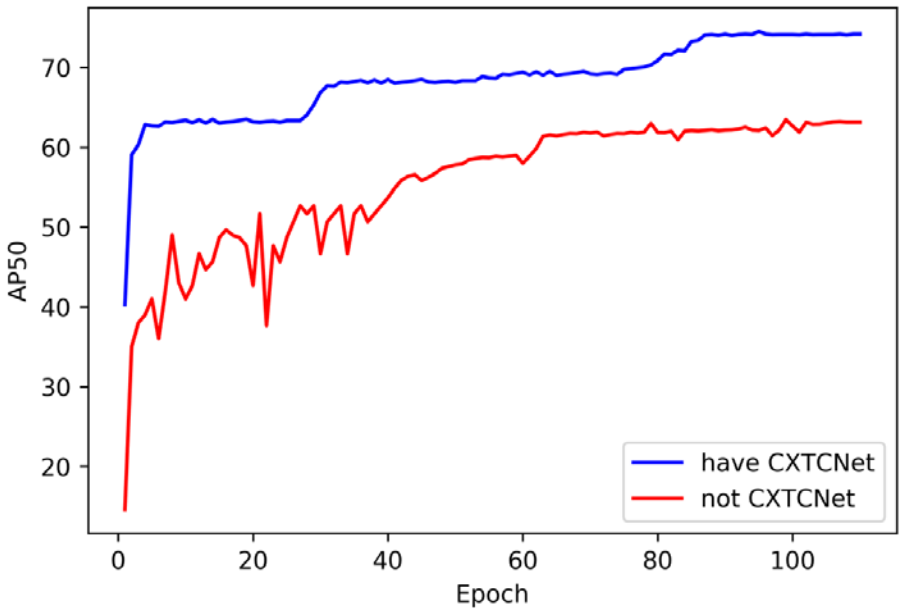


Fig. 12. When detecting whether the network uses CXTCNet, the change curve of AP50 during the training process, the blue is the change curve of using CXTCNet to screen in advance, and the red is the change curve of direct detection.

5.2.4 Comparison with other detection algorithms

The proposed chest X-ray TB detection algorithm is also compared with various mainstream target detection algorithms, including Faster R-CNN, SSD, RetinaNet, etc. The backbone of these detection algorithms is pre-training using ImageNet. **Table 7** shows the results of comparing our proposed algorithm with different detection algorithms. From **Table 7**, it can be seen that among these mainstream detectors, the detection of TB areas in chest X-ray images is poor. FOCS has the lowest AP50 of 46.6%, RetinaNet of 52.1%, and SSD of 52.3%, while Faster R-CNN has the highest AP50 of 57.3%. From the performance results of the detection, the performance of these detection algorithms is significantly lower than that of our proposed detection method.

Table 7. Comparison results between CXTDNet and multiple detection algorithms

Method	AP(%)	AP50(%)
Faster R-CNN	22.7	57.3
SSD	22.6	52.3
RetinaNet	22.2	52.1
FCOS	18.9	46.6
Our Method	**44.90**	**73.28**

5.2.5 Comparison with radiologists

Finally, to validate the effectiveness of our proposed algorithm based on attention mechanism and Sparse R-CNN for chest X-ray TB area detection in real-world scenarios, we tested it on our established dataset TBX304 and compared it with the diagnostic results of several radiologists in Shaanxi provincial tuberculosis control hospital. The experiments show that the accuracy of our proposed detection algorithm on this dataset is no less than that of professional physicians. **Fig. 13** shows the detection results of our proposed algorithm on TBX304.

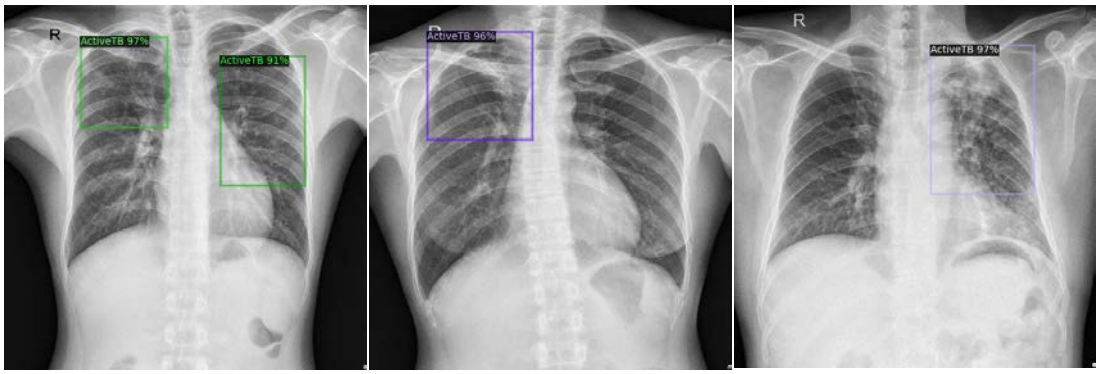


Fig. 13. The detection results of our proposed algorithm for TBX304 are shown here. The marked boxes in the figure show the TB areas detected by the algorithm, with the corresponding TB types and the calculated confidence.

6. Conclusion

Early diagnosis is essential for the treating and preventing tuberculosis, a major infectious disease. Inspired by the rapid development of computer-aided diagnosis systems and deep learning, we use CLAHE to preprocess the chest radiograph data to improve the contrast between key features of pulmonary tuberculosis and the background. Then, we propose a

tuberculosis detection algorithm using attention mechanism and Sparse R-CNN for detecting tuberculosis areas on chest X-ray images. The algorithm improves the reliability and accuracy of the whole algorithm by designing two networks, the chest X-ray TB classification network (CXTCNet) and the chest X-ray TB area detection network (CXTDNet). We design a channel attention module (CAM) in the classification network to enable the deep learning network to extract more helpful information about the image features. In the detection network, we designed CXTDNet based on the sparse object detection algorithm Sparse R-CNN, using a fixed set of learnable suggestion frames and learnable suggestion features for classification and localization. By combining CXTCNet and CXTDNet, we achieve the purpose of accurately locating TB areas in chest X-ray images. Moreover, the algorithm will distinguish latent TB from active TB during detection process. To further advance the development of TB detection, we build a new TB dataset called TBX304. The final experimental results demonstrate that our proposed attention mechanism and Sparse R-CNN based chest X-ray TB area detection algorithm outperforms the established detectors on the dataset TBX11K, In the dataset we build TBX304 performs no worse than radiology professionals. Although the chest X-ray TB area detection algorithm proposed in this paper can accurately locate the TB area in the image, the AP of the latent TB is relatively low due to the small amount of data on latent TB. We will consider adding latent tuberculosis samples to the newly created dataset or borrowing the current methods of small sample learning to study this problem. At the same time, the tuberculosis detection model is relatively complex and not easy to deploy, so reducing the complexity of the model is also a problem that needs to be solved in the next step of this research.

References

[1] J. Chakaya, M. Khan, F. Ntoumi, E. Aklillu, R. Fatima, P. Mwaba, N. Kapata, S. Mfinanga, S. E. Hasnain, P. D. M. C. Katoto, A. N. H. Bulabula, N. A. Sam-Agudu, J. B. Nachega, S. Tiberi, T. D. McHugh, I. Abubakar, and A. Zumla, "Global Tuberculosis Report 2020–Reflections on the Global TB burden, treatment and prevention efforts," *International Journal of Infectious Diseases*, vol. 113, pp. S7-S12, 2021. Article (CrossRef Link)

[2] E. Harding, "WHO global progress report on tuberculosis elimination," *The Lancet Respiratory Medicine*, vol. 8, no. 1, p. 19, 2020. Article (CrossRef Link)

[3] P. Andersen, M. E. Munk, J. M. Pollock, and T. M. Doherty, "Specific immune-based diagnosis of tuberculosis," *The Lancet*, vol. 356, no. 9235, pp. 1099-1104, 2000. Article (CrossRef Link)

[4] A. Bekmurzayeva, M. Sypabekova, and D. Kanayeva, "Tuberculosis diagnosis using immunodominant, secreted antigens of Mycobacterium tuberculosis," *Tuberculosis*, vol. 93, no. 4, pp. 381-388, 2013. Article (CrossRef Link)

[5] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 577-590, 2014. Article (CrossRef Link)

[6] A. Konstantinos, "Testing for tuberculosis," *Australian Prescriber*, vol. 33, pp. 12-18, 2010. Article (CrossRef Link)

[7] C. Miller, K. Lonnroth, G. Sotgiu, and G. B. Migliori, "The long and winding road of chest radiography for tuberculosis detection," *European Respiratory Journal*, vol. 49, no. 5, p. 1700364, 2017. Article (CrossRef Link)

[8] M. R. A. Van Cleeff, L. E. Kivihya-Ndugga, H. Meme, J. A. Odhiambo, and P. R. Klatser, "The role and performance of chest X-ray for the diagnosis of tuberculosis: a cost-effectiveness analysis in Nairobi, Kenya," *BMC infectious diseases*, vol. 5, no. 1, pp. 1-9, 2005. Article (CrossRef Link)

[9] S. Candemir, and S. Antani, "A review on lung boundary detection in chest X-rays," *International journal of computer assisted radiology and surgery*, vol. 14, no. 4, pp. 563-576, 2019. Article (CrossRef Link)
[10] J. Yanase, and E. Triantaphyllou, "A systematic survey of computer-aided diagnosis in medicine: Past and present developments," *Expert Systems with Applications*, vol. 138, pp. 112821, 2019. Article (CrossRef Link)
[11] J. Dinnes, J. Deeks, H. Kunst, A. Gibson, E. Cummins, N. Waugh, F. Drobniewski, and A. Lalvani, "A systematic review of rapid diagnostic tests for the detection of tuberculosis infection," *Health Technology Assessment*, vol. 11, no. 3, pp. 1-196, 2007. Article (CrossRef Link)
[12] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14454-14463, 2021. Article (CrossRef Link)
[13] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y. Wang, P. Lu, and C. J. McDonald, "Automatic tuberculosis screening using chest radiographs," *IEEE transactions on medical imaging*, vol. 33, no. 2, pp. 233-245, 2014. Article (CrossRef Link)
[14] R. B. Jeyavathana, R. Balasubramanian, and A. Pandian, "An Efficient Feature Extraction Method for Tuberculosis detection using Chest Radiographs," *International Journal of Applied Environmental Sciences*, vol. 12, no. 2, pp. 227-240, 2017.
[15] K. Satheeshkumar, and A. N. J. Raj, "Developments in computer aided diagnosis used for Tuberculosis detection using chest radiography: A survey," *Journal of Engineering and Applied Sciences*, vol. 11, no. 9, pp. 5530-5539, 2006.
[16] R. Hooda, S. Sofat, S. Kaur, A. Mittal, and F. Meriaudeau, "Deep-learning: A potential method for tuberculosis detection using chest radiography," in *Proc. of 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 497-502, 2017. Article (CrossRef Link)
[17] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386-397, 2020. Article (CrossRef Link)
[18] S. Gao, M. M. Cheng, K. Zhao, X. Y. Zhang, M. H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652-662, 2021. Article (CrossRef Link)
[19] Z. Cai, and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6154-6162, 2018. Article (CrossRef Link)
[20] R. Girshick, "Fast r-cnn," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1440-1448, 2015. Article (CrossRef Link)
[21] E. J. Hwang, S. Park, K. N. Jin, J. I. Kim, S. Y. Choi, and J. H. Lee, "Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs," *Clinical Infectious Diseases*, vol. 69, no. 5, pp. 739-747, 2019. Article (CrossRef Link)
[22] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. B. Mahbub, M. A. Ayari, and M. E. Chowdhury, "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol 8, pp. 191586-191601, 2020. Article (CrossRef Link)
[23] M. Ayaz, F. Shaukat, and G. Raja, "Ensemble learning based automatic detection of tuberculosis in chest x-ray images using hybrid feature descriptors," *Physical and Engineering Sciences in Medicine*, vol. 44, no. 1, pp. 183-194, 2021. Article (CrossRef Link)
[24] S. Jaeger, S. Candemir, S. Antani, Y. X. J. Wáng, P. X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, pp. 475-477, 2014. Article (CrossRef Link)

[25] A. Chauhan, D. Chauhan, and C. Rout, "Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation," *PloS one*, vol. 9, no. 11, pp. e112980, 2014. Article (CrossRef Link)
[26] Y. Liu, Y. H. Wu, Y. Ban, H. Wang, and M. M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2643-2652, 2020. Article (CrossRef Link)
[27] H. M. Blumberg, and J. D. Ernst, "The challenge of latent TB infection," *Jama*, vol. 316, no. 9, pp. 931-933. 2016. Article (CrossRef Link)
[28] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision Graphics & Image Processing*, vol. 39, no. 3, pp. 355-368, 1987. Article (CrossRef Link)
[29] S. Sajeev, M. Bajger, and G. Lee, "Segmentation of breast masses in local dense background using adaptive clip limit-CLAHE," in *Proc. of 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-8, 2015. Article (CrossRef Link)
[30] J. C. M. dos Santos, G. A. Carrijo, C. D. F. dos Santos Cardoso, J. C. Ferreira, P. M. Sousa, and A. C. Patrocínio, "Fundus image quality enhancement for blood vessel detection via a neural network using CLAHE and Wiener filter," *Research on Biomedical Engineering*, vol. 36, pp. 107-119, 2020. Article (CrossRef Link)
[31] R. M. James, and A. Sunyoto, "Detection Of CT-Scan Lungs COVID-19 Image Using Convolutional Neural Network And CLAHE," in *Proc, of 2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 302-307, 2020. Article (CrossRef Link)
[32] 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. Article (CrossRef Link)
[33] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354-377, 2018. Article (CrossRef Link)
[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. Article (CrossRef Link)
[35] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4700-4708, 2017. Article (CrossRef Link)
[36] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011-2023, 2020. Article (CrossRef Link)
[37] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Proc. of 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, pp. 844-848, 2014. Article (CrossRef Link)
[38] W. Rawat, and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352-2449, 2017. Article (CrossRef Link)
[39] L. Perez, and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017. Article (CrohssRef Link)
[40] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. of the 26th International Conference on Neural Information Processing Systems*, Vol. 2, pp. 2553-2561, 2013. Article (CrossRef Link)
[41] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning system*, vol. 30, no. 11, pp. 3212-3232, 2019. Article (CrossRef Link)
[42] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788, 2016. Article (CrossRef Link)

[43] J. Redmon, and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7263-7271, 2017. Article (CrossRef Link)
[44] J. Redmon, and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:* 1804.02767, 2018. Article (CrossRef Link)
[45] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, "Ssd: Single shot multibox detector," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 21-37, 2016. Article (CrossRef Link)
[46] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020. Article (CrossRef Link)
[47] M. Tan, and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. of International Conference on Machine Learning (ICML)*, pp. 6105-6114, 2019. Article (CrossRef Link)
[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91-99, 2015. Article (CrossRef Link)
[49] N. Bodla, B. Singh, R. Chellappa, and L.S. Davis, "Soft-NMS--improving object detection with one line of code," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5561-5569, 2017. Article (CrossRef Link)
[50] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 213-229, 2020. Article (CrossRef Link)
[51] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117-2125, 2017. Article (CrossRef Link)
[52] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3111-3122, 2018. Article (CrossRef Link)
[53] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658-666, 2019. Article (CrossRef Link)
[54] S. Vajda, A. Karargyris, S. Jaeger, K. C. Santosh, S. Candemir, Z. Y. Xue, S Antaniet, and G. Thoma, "Feature selection for automatic tuberculosis screening in frontal chest radiographs," *Journal of medical systems*, vol. 42, no. 8, pp. 1-11, 2018. Article (CrossRef Link)
[55] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization," *Scientific reports*, vol. 9, no. 1, pp. 1-9, 2019. Article (CrossRef Link)
[56] E. Tasci, C. Uluturk, and A. Ugur, "A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection," *Neural Computing and Applications*, vol. 33, no. 22, pp. 15541-15555, 2021. Article (CrossRef Link)
[57] R. Guo, K. Passi, and C. K. Jain, "Tuberculosis diagnostics and localization in chest X-rays via deep learning models," *Frontiers in Artificial Intelligence*, vol. 3, 2020. Article (CrossRef Link)
[58] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 740-755, 2014. Article (CrossRef Link)

Xuebin Xu received Ph.D. degree in the department of computer science from Xi'an Jiaotong Univeristy, PR China in 2010. He has worked as a research associate professor of University of Florida, USA. He is currently as a research scientist of Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing. His research interests include machine learning, bioinformatics, biomedical image processing and information fusion. He has published over 100 papers in top international journals and conferences.

Jiada Zhang was born in Chaozhou, China in 1996. He is currently pursuing a master's degree in computer technology at Xi 'an University of Posts and Telecommunications. His research interests include image recognition and object detection.

Xiaorui Cheng was born in Baoji, China in 1997. She is currently pursuing a master's degree in computer Science and technology at Xi 'an University of Posts and Telecommunications. Her research interests include medical big data and eeg processing.

Longbin Lu received his ph. D. degree in Control Science and Engineering from Xi 'an Jiaotong University in 2018. He is currently a lecturer at the School of Computer Science, Xi 'an University of Posts and Telecommunications. His research interests include biometric recognition and artificial intelligence.

Yuqing Zhao was born in 1995 in Weinan, China. She is currently pursuing a master's degree in computer technology at Xi 'an University of Posts and Telecommunications. Her research interests include image recognition and image segmentation.

Zongyu Xu was born in Xi 'an, China in 1997. He is currently pursuing a master's degree in computer technology at Xi 'an University of Posts and Telecommunications. His research interests include intelligent transportation and deep learning.

Zhuangzhuang Gu was born in Kaifeng, China in 1998. He is currently pursuing a PhD in Computer Science at the University of South Carolina. His research interests include machine learning and deep learning.