applied sciences

MDPI

Article

An Improved Algorithm for Detecting Pneumonia Based on YOLOv3

Shangjie Yao [1], Yaowu Chen [2,*], Xiang Tian [3], Rongxin Jiang [4] and Shuhao Ma [5]

1 Institute of Advanced Digital Technology and Instrumentation, Zhejiang University, Zhejiang 310027, China; 11415009@zju.edu.cn
2 The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China
3 Zhejiang Provincial Key Laboratory for Network Multimedia Technologies, Hangzhou 310027, China; tianx@zju.edu.cn
4 Zhejiang University Embedded System Engineering Research Center, Ministry of Education of China, Hangzhou 310027 China; rongxinj@zju.edu.cn
5 The Institute of Information Science and Technology Instrumentation, Dalian Maritime University, Dalian 116026, China; mashuhao1789@126.com
* Correspondence: cyw@mail.bme.zju.edu.cn

Received: 19 January 2020; Accepted: 28 February 2020; Published: 6 March 2020

Abstract: Pneumonia is a disease that develops rapidly and seriously threatens the survival and health of human beings. At present, the computer-aided diagnosis (CAD) of pneumonia is mostly based on binary classification algorithms that cannot provide doctors with location information. To solve this problem, this study proposes an end-to-end highly efficient algorithm for the detection of pneumonia based on a convolutional neural network—Pneumonia Yolo (PYolo). This algorithm is an improved version of the Yolov3 algorithm for X-ray image data of the lungs. Dilated convolution and an attention mechanism are used to improve the detection results of pneumonia lesions. In addition, double K-means is used to generate an anchor box to improve the localization accuracy. The algorithm obtained 46.84 mean average precision (mAP) on the X-ray image dataset provided by the Radiological Society of North America (RSNA), surpassing other detection algorithms. Thus, this study proposes an improved algorithm that can provide doctors with location information on lesions for the detection of pneumonia.

Keywords: convolutional neural network; pneumonia detection; medical image

1. Introduction

In recent years, the number of people suffering from pneumonia in the world has increased year by year. In particular, the incidence of pneumonia in infants has increased significantly, which seriously threatens the survival and health of human beings [1]. At present, the essence of most computer-aided diagnosis (CAD) system algorithms for the lungs is image classification. Although this kind of algorithm has the advantages of simple implementation and high accuracy, the output results lack the location information of lesion tissue, so it cannot provide more valuable reference for doctors. The algorithm proposed in this study benefits from the accurate labeling of datasets [2] and the rapid development of a convolutional neural network (CNN)-based objection detection algorithm, which enables it to identify pneumonia and locate pneumonia tissue at the same time, so it can provide more reference information for doctors.

The CNN is a kind of artificial neural network with a deep structure and convolution calculation. The CNN has the ability of representation learning, which can capture the spatial local correlation of the input data through convolution operation and obtain the translation invariance of the input

data through down-sampling operation (see Figure 1). Nowadays, the CNN has become the focus of research in many scientific fields, particularly in the field of computer vision. The advantages of the CNN are that complex pre-processing of images is not required and they can use original images as direct input; thus, they have a wide range of applications.

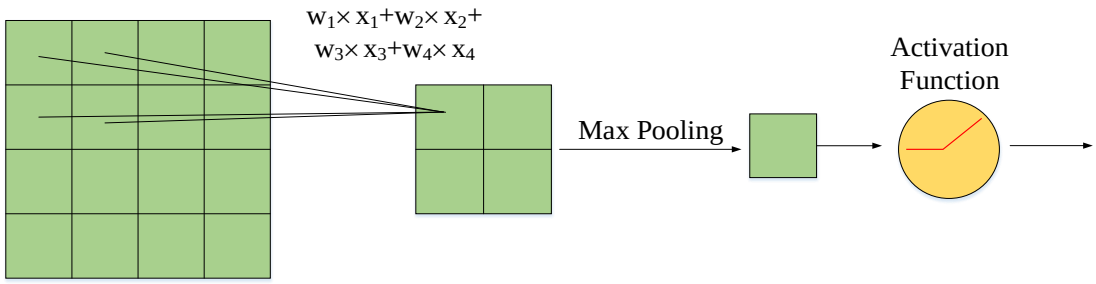


Figure 1. Basic unit of convolutional neural network. The Max Pooling is responsible for obtaining translation and rotation invariance, and the Activation Function is responsible for increasing the degree of nonlinearity.

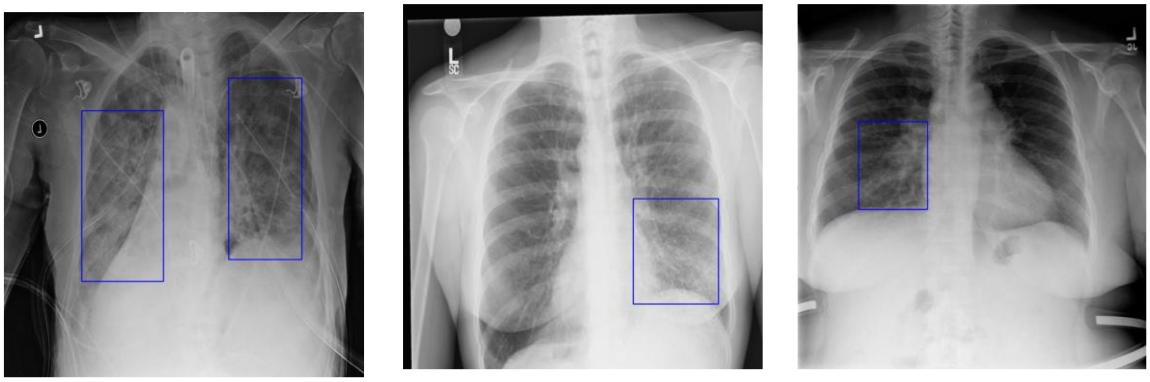
Girshick et al. [3] first proposed to train the CNN to extract features using back propagation and used support vector machine (SVM) as a classifier to construct a region-based CNN (RCNN) object detection algorithm. However, the RCNN needs to pre-select regions where objects may exist using selective search methods and separate the extraction feature phase from the classification phase. Such an algorithm is not end-to-end, it is difficult to implement, and it has low computational efficiency. Girshick [4] and Ren et al. [5] proposed an end-to-end detection algorithm that integrates the selection of object regions, extraction features, and classification using an extraction feature network and region proposal network (RPN). However, this algorithm is a two-stage detection algorithm that has high hardware requirements and has difficulty achieving real-time detection in the training and testing phases. Redmon et al. [6] introduces an end-to-end real-time object detection algorithm called Yolo that uses the CNN to perform the extraction feature, classification, and localization of the object. It is a one-stage detection algorithm that has a high detection speed but unsatisfactory performance on object localization. The Yolov2 algorithm was proposed by Redmon and Farhadi [7]. This algorithm uses K-means clustering to cluster anchor scales as prior knowledge in many datasets to improve object localization. In addition, random scaling has proposed for use to enhance the generalization of the algorithm to different scales. The Yolov3 algorithm, subsequently proposed by Redmon and Farhadi [8], uses Feature Pyramid Net (FPN) [9] to improve performance with respect to missed detection of small objects in Yolov2.

The above are the study of CNN-based object detection algorithms. Next, we will learn about the CAD system.

At present, most CAD systems for the lung rely on image classification algorithms. This kind of system takes a whole image as input, extracts the feature using a feature extractor, and finally obtains a predictive label of the image from the classifier. This kind of CAD system cannot provide the doctors with accurate location information on lesions; therefore, its usefulness is limited. For example, in a study by Varshni et al. [10], the CNN was used to extract features, and an SVM was used as a classifier to detect pneumonia from an input image. In a study by Setio et al. [11], multiple pulmonary nodule detection algorithms were proposed and combined with the CNN to construct a CAD system that was shown to achieve detection sensitivities of 85.4% and 90.1% at 1 and 4 false positives per scan on lung image database consortium and image database resource initiative (LIDC-IDRI) [12], which is a small-scale dataset. In a study by Rajpurkar et al. [13], a large-scale pneumonia detection dataset was proposed, and a network with 121 layers of convolution was proposed to detect pneumonia. The essence of these studies is the classification of X-ray images; therefore, the CAD systems constructed can only provide category information, not location information as a reference. However, it is difficult for the naked eye to distinguish pneumonia lesions from normal tissues in X-ray images and therefore

it is important to construct CAD systems that can provide pneumonia lesion location information. There are two difficulties encountered in constructing such a CAD system: 1) judging the presence of objects in the image and 2) accurately locating objects. The currently available CNN-based object detection algorithms can be used to construct end-to-end CAD systems to provide location information, but the algorithms need to be improved to fit X-ray image datasets.

In the study, Yolov3 was improved by analyzing the advantages and disadvantages of the existing algorithms and combining the characteristics of the pneumonia dataset (see Figure 2). Small differences in the characteristics of lesion and non-lesion were believed to increase the difficulty of detecting lesion, but increasing the perception field of the algorithm so that it could use the global information in the image was expected to increase the recognition ability. Therefore, multi-branch dilated convolution [14,15] was added to Yolov3. In the Yolov3 object detection algorithm, multiple down-sampling operations result in the loss of semantic information and spatial information, making differences between lesion and non-lesion smaller in the feature space and improving the recognition ability of the algorithm. Although the fusion of low-level and high-level features can solve this problem to a certain extent, the use of attention mechanism to suppress the output of inaccurate semantic information in low-level features further improves the performance of the algorithm. The use of double K-means enables the algorithm to generate anchor boxes of different scales for different input images.



(**a**) (**b**) (**c**)

Figure 2. Experimental data used in this study. The area within the blue box shows a pneumonia lesion. In images (**a**) and (**b**), there is no significant difference between pneumonia lesion and non-lesion. It can be seen from image (**c**) that the features of the left lesion and the right non-lesion are similar.

This study describes the propose detection algorithm, i.e., Pneumonia Yolo (PYolo) for pneumonia. PYolo uses multi-branch dilated convolution to increase the perception field and attention mechanism to suppress the output of inaccurate semantic information in low-level features and enhance the ability of the algorithm to detect a lesion. In addition, this paper proposes the use of double K-means to generate anchor boxes to improve the localization accuracy.

2. Materials and Methods

2.1. Materials

Earlier research on attention focuses on the analysis of brain imaging, which will not be introduced in detail in this chapter. At a time when deep learning grows vigorously, it is prominent to construct CNN with the attention mechanism. On the one hand, the neural network can learn the attention mechanism autonomously; on the other hand, the attention mechanism can in turn help us understand the world presented by neural network. In recent years, most of the research on the combining of deep learning and visual attention [16–18] focuses on using masks to form the attention mechanism. The principle of using the mask as the attention mechanism lies in the extraction of key features from the image using the weights predicted by the neural network. Through learning and training,

the neural network can learn the areas that need attention in each new image. This is the purpose of the attention mechanism in deep learning.

In the field of semantic segmentation, the architecture of the neural network generally adopts the fully convolutional network (FCN) [19]. The FCN, like the traditional CNN, first performs convolution operations on images and then performs pooling operations to reduce the image size and increase the receptive field. However, since semantic segmentation is a pixel-wise output, the smaller image size obtained after the pooling operation is up-sampling to the original image size for prediction (up-sampling is generally made by bilinear interpolation). In this regard, there are two key operations in the segmentation algorithm: one is pooling, to reduce the image size and increase the receptive field; and the other is up-sampling, to increase the image size. In the process of reducing the size of the image, FCN loses some of the spatial and semantic information in images, which is not conducive to segmentation. Therefore, a convolution operation that can obtain a large receptive field without pooling is introduced, i.e., dilated convolution [20]. Dilated convolution introduces a parameter called dilation rate, which defines the distance of sampling by the kernel. The larger the dilation rate, the larger the sampling distance, and the larger the receptive field of the kernel.

2.2. Methods

In the following sections, we describe the detection process of PYolo in general and then introduce each of the three algorithm improvements proposed in this paper: 1) location pre-processing, 2) MaskFPN, and 3) dilated convolution.

2.2.1. PYolo Detection Process

As shown in Figure 3a, in the location pre-processing, PYolo uses double K-means to produce the anchor box of a lesion. As shown in Figure 3b, PYolo uses DarkNet53 to extract features, uses MaskFPN to fuse features of different levels, and uses a multi-branch convolution module to obtain multi-perception field information. Unlike Yolov3, PYolo only detects the features of the module output. The input image size of DarkNet53 is 416 × 416 pixels, and the output features are {F1, F2, F3} with the sizes of {13 × 13, 26 × 26, 52 × 52}, respectively. In the experiment, the input image was scaled to 416 × 416 pixels in the pre-processing stage. The difference between MaskFPN and FPN is that MaskFPN uses the information of high-level feature as prior knowledge to generate a weight map, and then multiplies the weight map with low-level features linearly to suppress the output of inaccurate semantic information of low-level features. By contrast, FPN directly combines high-level features and low-level features, directly overcoming the problem of inaccurate semantic information in low-level features.

In an object detection algorithm, the ratios of positive and negative samples are critical to the performance of the algorithm. As shown in Figure 4, like Yolov3, PYolo corresponds to the feature points by dividing the image into grid cells. In the training phase, the real bounding box is mapped to the corresponding coordinates on the feature map by dividing by the stride; in the detection phase, the predicted bounding box on the feature map is mapped to the corresponding coordinates on the original image by multiplying the stride. The dimensions of the output features of PYolo are [S, S, A * (B + Conf + Cls)]. S × S is the number of grid cells; B is the predicted bounding box; Conf is the confidence level of the output object; Cls is the class of the dataset; and A is the number of scales for each anchor. With respect to the selection of positive and negative samples, anchors with the Intersection over Union (IOU) with the ground-truth bounding boxes were used for evaluation. Anchors that have IOU with any ground-truth box greater than 0.5 were included as training samples. The center points of the anchor and ground-truth bounding boxes that fall on the same grid were designated as positive samples, and other anchors were designated as negative samples.

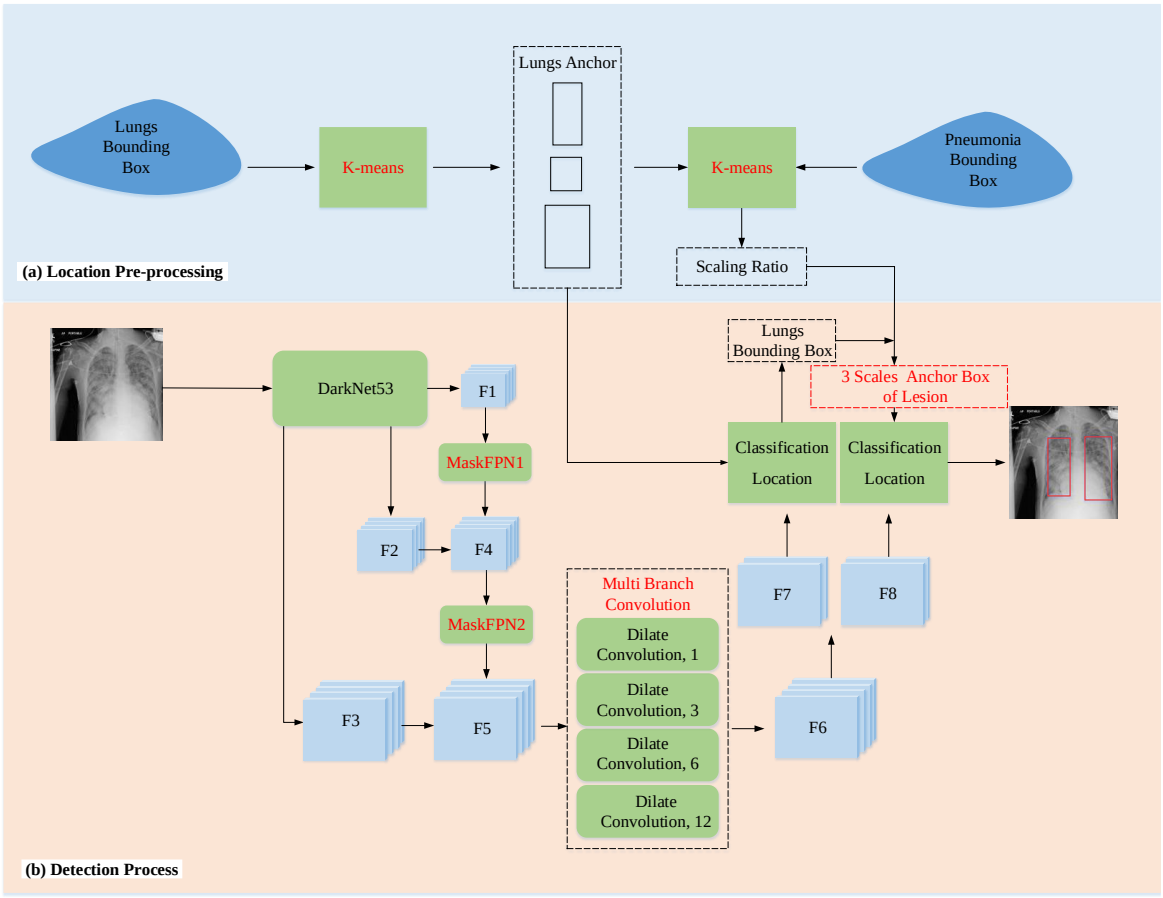


Figure 3. Structure of Pneumonia Yolo (PYolo). The highlighted red text indicates improvements based on Yolov3 in this study. In the multi-branch convolution module, 1, 3, 6, and 12 represent different dilation rates: (**a**) location pre-processing; (**b**) detection process.

There is still a problem of imbalance between positive and negative samples in the screened sample set. To overcome the problem of imbalance between positive and negative samples [21], a hyper-parameter $\lambda = 200$ is introduced in the loss function to strengthen the learning intensity for negative samples and accelerate the speed of the convergence of the model. The localization loss function is different from the function in Yolov3. Smooth L1 loss was adopted as the localization loss function as it has a higher level of smoothness compared to others. The loss functions of the model are as follows:

$$L_{loc} = \sum_{i \in Pos}^{N} \sum_{m \in \{x,y,w,h\}} \text{smooth}_{\text{L1}}(g_m - p_m) \tag{1}$$

$$L_{cls} = -\sum_{i \in Pos} C_i \log(X_i) \tag{2}$$

$$L_{pos} = -\sum_{i \in Pos} M_i \log(Y_i) \tag{3}$$

$$L_{neg} = -\sum_{i \in Neg} M_i \log(Y_i) \tag{4}$$

$$L_{total} = L_{loc} + L_{cls} + L_{pos} + \lambda L_{neg} \tag{5}$$

where L_{loc}, L_{cls}, L_{pos}, and L_{neg} represent localization loss, classification loss, positive sample loss, and negative sample loss, respectively, and g, p, C, X, M, and Y refer to the actual coordinates, predicted coordinates, probability of the actual class, and probability of the predicted class, actual set of positive

and negative samples, and predicted set of positive and negative samples, respectively. Equation (5) is the overall loss function of the algorithm.

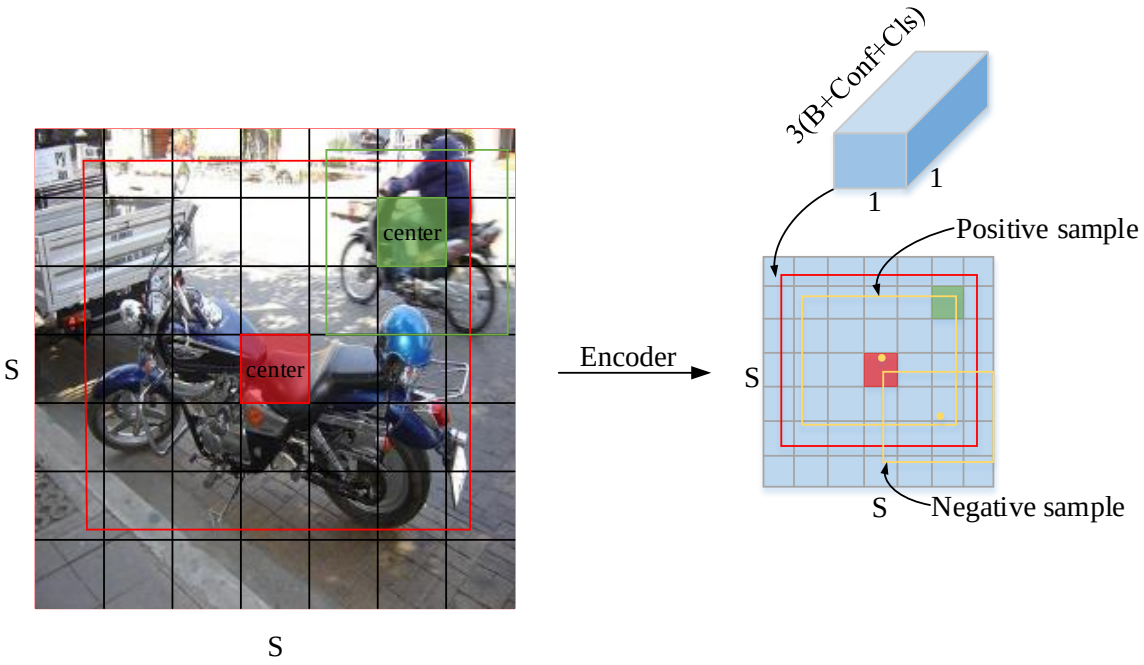


Figure 4. Mapping relationship between the image and feature. In the left image, the red and green boxes are ground-truth boxes. In the right feature, the red box is a ground-truth box, and the yellow box is an anchor. The center points of the anchor and ground-truth bounding boxes that fall on the same grid cell were designated as positive samples, and other anchors were designated negative samples. If the Intersection over Union (IOU) of the anchor and ground truth is greater than the threshold but the center points do not fall on the same grid cell, then the anchor it is regarded as a negative sample.

2.2.2. Double K-Means

The anchor box is a preset bounding box size. Regression of the anchor box helps to improve the localization accuracy of the algorithm. K-means is used in Yolov3 to generate the anchor box. In the algorithm proposed in this study, double K-means is used to generate the anchor box for lesions proposed in a specific method for the pneumonia dataset. The method consists of two phases. In the first phase, K-means is used to generate the lung anchor box for the algorithm to locate the lung. In this study, the lung anchor box with three scales of {[78, 136], [129, 207], [163, 256]} were generated. In the second phase, K-means is used again to generate one scale ratio for each lung anchor box; therefore, in PYolo, the anchor box of three scales was obtained for lesions. Figure 5 shows the steps involved in generating the scaling ratio, where the lesion-bounding box and lung-bounding box are clustered into the three clusters shown in the Figure 5 through K-means clustering. The mean IOU for the two kinds of bounding box in each cluster is calculated to obtain the scale ratio. In this study, the scale ratio of the anchor box with the size of {[78, 136], [129, 207], [163, 256]} was {0.7, 0.7, 0.5}.

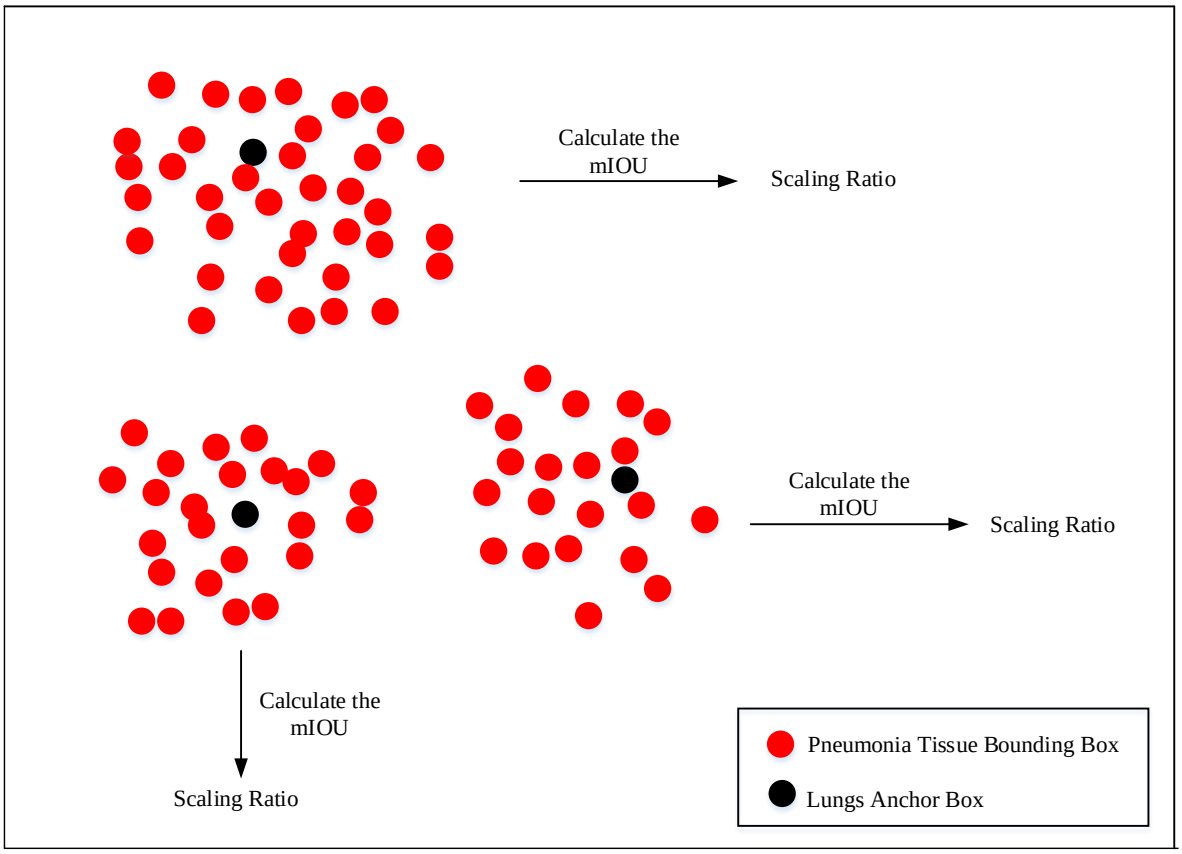


Figure 5. Schematic of scaling ratio generation. The black dots represent the cluster center; mIOU is the average value of the IOU of the black points and all the red points in each cluster, and this value is taken as the scaling ratio.

2.2.3. MaskFPN for Suppression of Information

MaskFPN is an improvement aspect of the attention mechanism. In the study by Hu et al. [22], the channel-wise weighting obtained by global pooling was too coarse. MaskFPN proposed in this paper assigns a weight to each pixel of the low-level feature map by generating pixel-wise weights to suppress the output of inaccurate semantic information in low-level features. As shown in Figure 6, MaskFPN performs linear multiplication on a set of feature maps C = {C_1, C_2, C_3, C_4} and a set of weight maps W = {W_1, W_2, W_3, W_4}. The weight value of each pixel in the weight map is in the interval [0, 1], where c_i and w_i have the same dimensions, i∈{1, 2, 3, 4}. For a pixel with a small weight value in the weight map, the intensity of the suppression of information at its corresponding position in the low-level feature is strong; and for a pixel with a large weight value, the intensity of suppression of information at its corresponding position in the low-level feature is low.

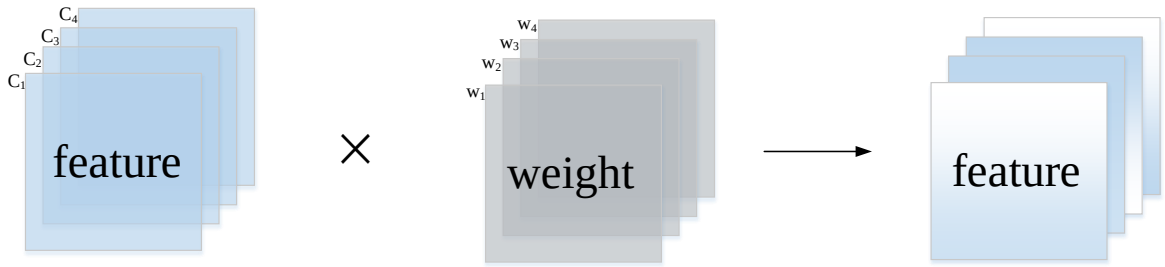


Figure 6. MaskFPN suppression of feature information. The white part of the feature map represents the semantic information filtered by MaskFPN.

In PYolo, each MaskFPN consists of two convolutional layers, a batch normalization, a leaky relu activation function, and a squeezing function. Figure 7 shows the specific flow of MaskFPN. The feature map F1 is converted by the first convolution layer, batch normalization, leaky relu, and the second convolution layer, and generates feature and weight maps. Then, each weight value in the weight map is converted into a value in the [0, 1] range through convolution and the squeezing function. The output after linear multiplication of the weight map and feature map F2 is combined with the high-level features to generate F4.

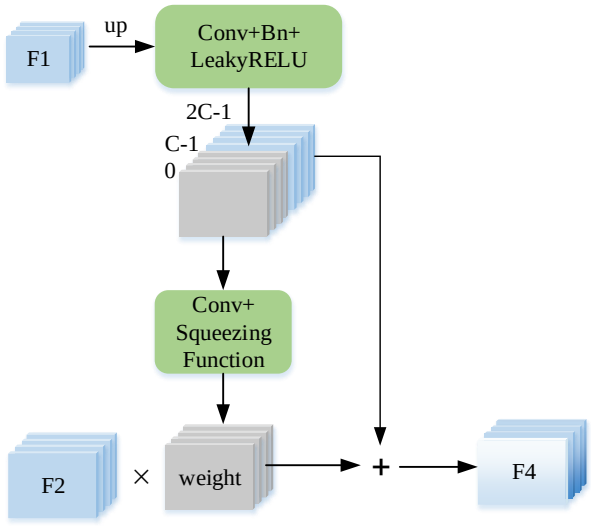


Figure 7. Process of MaskFPN. The information output from the feature map F1 after the convolution operation is dilated by a factor of 2 channel-wise. The information from the [0 ~ C-1] channel generates the weight map by the squeezing function. The information from channel [C ~ 2C-1] is used for feature fusion.

Table 1 presents the specific parameter settings of MaskFPN. It can be seen that the combination of the kernel size, stride, padding, and dilation rate in MaskFPN does not reduce the resolution of the feature map. It only transforms feature information in the dimension of the channel.

Table 1. MaskFPN parameter settings.

Module	Component	Kernel Size	Padding	Stride	Input Channels	Output Channels	Dilation Rate
MaskFPN 1	Conv Layer1 BatchNorm2d Leaky ReLU	3	1	1	512 1024	1024	1
	Conv Layer2 Squeezing Function	3	1	1	512	255	1
MaskFPN 2	Conv Layer1 BatchNorm2d Leaky ReLU	3	1	1	256 512	512	1
	Conv Layer2 Squeezing Function	3	1	1	256	255	1

2.2.4. Dilated Convolution for Capturing Information in Multiple Receptive Fields

Humans usually rely on relevant feature information to guess the objects that are difficult to be recognized. In the dataset used in this study, there is no significant difference between pneumonia lesion and non-lesion, so two parallel dilated convolution layers were introduced for the PYolo algorithm to capture global information to increase the prediction ability of the algorithm. Dilated convolution expands the perception field of the kernel by inserting 0 in it. The advantage of the kernel is that it can obtain a larger range of view without down-sampling. For the case where the features of the

pneumonia lesion are not obvious, the features are easily lost during the down-sampling process, which leads to a decrease in the accuracy of the algorithm. The algorithm uses dilated convolution to increase its own perception field and avoid the loss of semantic and spatial information caused by down-sampling. The dilation rate is usually selected based on the principle of not reducing the resolution of the feature map and multiple perception field. A kernel with a smaller perception field can obtain local information of features, while the kernel with a larger perception field can obtain the global information of feature [23]. However, a too large dilation rate of the kernel can degrade the performance of the kernel. The reason of this is that the kernel would have a too large receptive field to capture the local dependencies in the image, and too many dilated convolution branches will cause an increase in the computation of dot products, affecting the forward propagation speed of the algorithm. The proposed algorithm of the study uses four convolution branches with different dilation rates. Table 2 shows the parameter settings in the convolution layer:

Table 2. Parameter settings in the dilated convolution layer.

Kernel Size	Dilation Rate	Padding	Stride	Input Channels	Output Channels
3	1	1	1	255	255
3	3	3	1	255	255
3	6	6	1	255	255
3	12	12	1	255	255

3. Results

3.1. Experimental Data

The dataset selected for the experiments in this study consists of a total of 6000 chest X-ray images provided by the Radiological Society of North America (RSNA). The 600 images in the dataset were randomly divided into test set, and the remaining 5400 images were used as training set. The training set was augmented to 10,800 images and the augment technique was only horizontal flip. The ratios of images with pneumonia lesions in the 10,800 training set and 600 test set were 0.65 and 0.70, respectively. Each input image was an original single-channel grayscale image, which was converted into a three-channel image with a resolution of 1024 × 1024 pixels during the image pre-processing phase. The bounding box of the lesion and lungs was labeled as (x, y, w, h), where x and y are the coordinates of the upper left corner of the object, and w and h are the length and width of the object. The lung bounding box was manually marked by the author of this paper.

3.2. Experimental Settings

In this study, experiments were ran with PyTorch 0.4, which was developed by Facebook in the United States. In order to increase the generalization ability of the model, the algorithm used pre-trained weights in Yolov3. The initial learning was 0.005, the learning rate schedule was polynomial decay, the momentum was set to 0.0005, and the optimizer adopted SGD [24–26], the weight decay was 0.0005, and the activation function used the leaky relu function [27]. BatchNorm [28] was used to prevent gradient descent during the training phase and accelerate the convergence of the model. The batch size was set to six.

3.3. Performance Indicators

The accuracy index of this study was defined as the average precision (AP) calculated from precision and recall.

Precision indicates the percentage of actual positive samples out of predicted positive samples. There are two sources of predicted positive samples: one is the TP number of positive samples predicted as positive samples; the other is the FP number of negative samples predicted as positive samples. Therefore, precision is calculated as P = TP/(TF + FP).

Recall refers to the percentage of total positive samples in the sample that are predicted correctly. The sample set includes: the TP number of predicted positive samples, and the FN number of predicted negative samples. Recall is calculated as R = TP/(TP + FN).

3.4. Ablation Experiment

In order to verify the effectiveness of the three improvements proposed in this study, the performance of Yolov3 was used as the baseline, and the three improvements were combined with Yolov3 in the ablation experiment conducted.

3.4.1. Double K-Means

Figure 8a,b show the effects of K-means and double K-means on detection effectiveness, where the calculation formula is the IOU between the predicted bounding box and the real bounding box obtained through non-maximum suppression NMS and then divided by the number of predicted bounding boxes. It can be seen from (a) that the value of IOU increases gradually with the increase of iterations. However, at the beginning of the iteration, the IOU of the lesion bounding box predicted by Yolov3 was very low compared to the real bounding box. According to the analysis of this study, in the early phase of the training, due to the inaccurate pulmonary bounding box predicted by Yolov3, the anchor box of the lesion was not accurate enough; therefore, the IOU value of predicated lesion bounding box was low. It can be seen from (b) that, during the training phase, the loss value of Yolov3 in locating the lesion is higher than that in the lungs because the lung features are relatively obvious and it is easier for PYolo to locate them, while the pneumonia lesions were more similar to the normal lung, which increased the learning difficulty for PYolo.

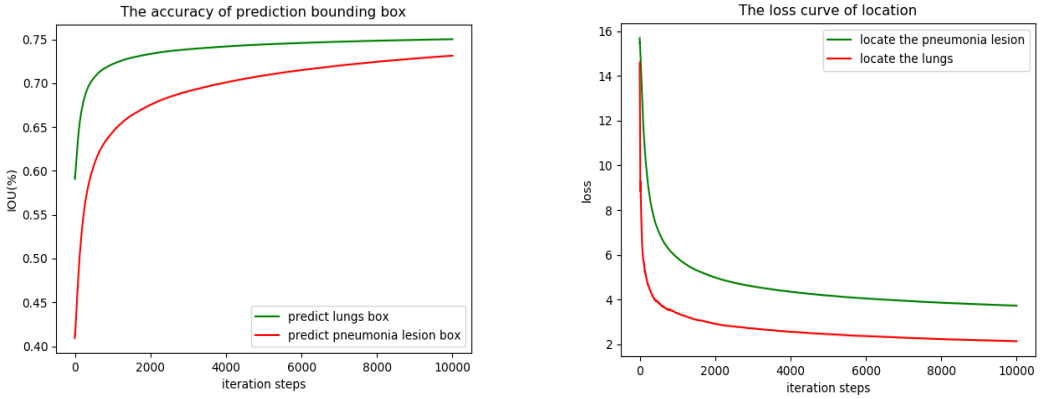


Figure 8. Performance of double K-means.

In order to improve the precision of the experiment, the test results under three different IOU thresholds were tested in the experiment, and the mAP in Table 3 is the average value of the test results with threshold values of {0.4, 0.5, 0.6}. '@' means that the accuracy of the algorithm is tested with the IOU threshold set. It can be seen that the AP value of Yolov3 obtained using the double K-means is higher than that obtained using K-means for each of the three thresholds. The size of the anchor box generated by double K-means varies with the input data, with stronger flexibility. However, as can be seen from Figure 8, the performance of double K-means proposed in this paper depends on the accuracy of the algorithm used in locating the lung.

Table 3. Comparison of double K-means with K-means (%).

Clustering	AP@0.4	AP@0.5	AP@0.6	mAP
K-means	57.91	43.01	29.22	43.40
Double K-means	59.52	45.27	31.08	45.29

3.4.2. MaskFPN

Figure 9 and Table 4 show the performance of sigmoid, tanh, and softmax as the squeezing function for the last layer of MaskFPN. It can be seen from Table 4 that MaskFPN using a sigmoid function improved the detection performance of Yolov3 to the highest level in comparison with FPN, but the detection performance of MaskFPN using a softmax function decreased in comparison with FPN. It can also be seen from Figure 9 that the overall convergence effect of the Yolov3 algorithm using the softmax function was not as satisfactory as that with the tanh and sigmoid functions. According to the analysis of this study, each probability score of MaskFPN output is relatively low due to the use of the softmax function in high-dimensional space, so the mask value in the weight map is small, which greatly reduces the information of low-level features, resulting in insufficient feature expression ability after fusion.

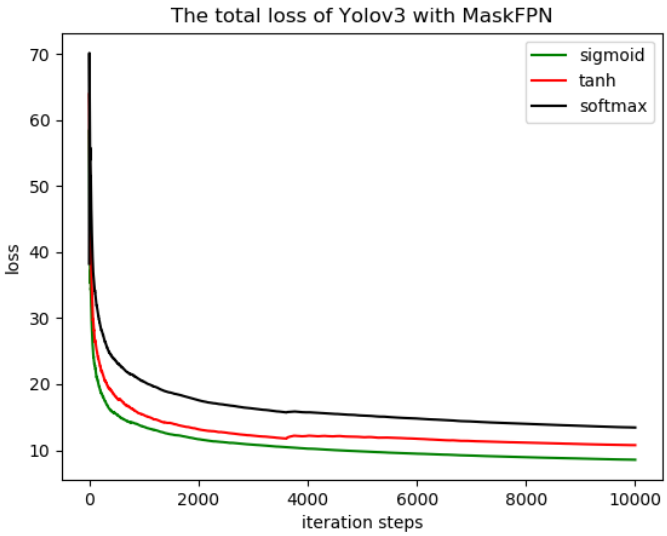


Figure 9. Performance of MaskFPN in the training phase.

Table 4. Comparison of MaskFPN and FPN (%).

Generate Anchor Box	Fuse Feature Approach	AP@0.4	AP@0.5	AP@0.6	mAP
Double K-means	FPN	57.91	43.01	29.22	43.40
Double K-means	MaskFPN_sigmoid	59.34	43.97	29.87	44.39
Double K-means	MaskFPN_tanh	58.62	43.59	29.65	43.95
Double K-means	MaskFPN_softmax	54.01	40.53	26.59	40.38

3.4.3. Dilated Convolution

Table 5 shows the detection performance of Yolov3 with dilated convolution branches. As seen from the data in the table, the detection performance of the algorithm gradually improves as the expansion rate increases. The mAP of Yolov3 increased by 2.20% when the dilation rate was {1, 3, 6, 12}, compared with a dilation rate of 1. The increase in convolution branches with different dilation rates means that the algorithm can obtain more information on the perception field. Because of hardware limitations, it was not possible in this study to continue to explore more dilated convolution in the detection algorithm. It has been suggested [15], however, that excessive dilated convolution prevents

capture of the local spatial image correlation; therefore, the kernel size degenerates to a 1×1 size, which prevents the continued improvement of the detection performance of the algorithm.

Table 5. Dilated convolution performance (%).

Dilation Rate	AP@0.4	AP@0.5	AP@0.6	mAP
1	57.91	43.01	29.22	43.40
1,3	58.64	44.37	30.13	44.38
1,3,6	59.36	44.93	30.72	45.00
1,3,6,12	60.56	45.23	31.02	45.60

3.5. Comparison of Detection Performance of Different Algorithms

Table 6 shows the effection of double K-means, MaskFPN and dilated convolution for Yolov3 with $\lambda = 200$. It can be seen from the Table 6 that when double K-means, MaskFPN and dilated convolution are used alone or together for Darknet53, the algorithm's mAP is improved. The parameter settings in MaskFPN and dilated convolution are the same as in Tables 1 and 2. We also evaluated the effects of different hyper-parameter values on controlling negative samples in the training phase, which is important in overcoming the problem of imbalance between positive and negative samples. Table 7 shows the detection accuracy of PYolo for different hyper-parameter values.

Table 6. Impact of improvements on PYolo (%).

Encoder	FPN	Double K-means	MaskFPN	Dilated Convolution	mAP
DarkNet53	√				43.40(Yolov3)
DarkNet53	√	√			44.82
DarkNet53		√	√		45.74
DarkNet53		√	√	√	46.84(PYolo)

Table 7. Effect of hyper-parameter value on precision and convergence speed of PYolo (%).

Hyper-Parameter	AP@0.4	AP@0.5	AP@0.6	Iterations Steps
$\lambda = 50$	45.17	36.62	24.16	200k
$\lambda = 100$	48.53	39.59	25.39	140k
$\lambda = 150$	53.15	42.87	26.65	115k
$\lambda = 200$	64.16	44.71	31.63	90k
$\lambda = 250$	56.67	42.43	24.73	90k

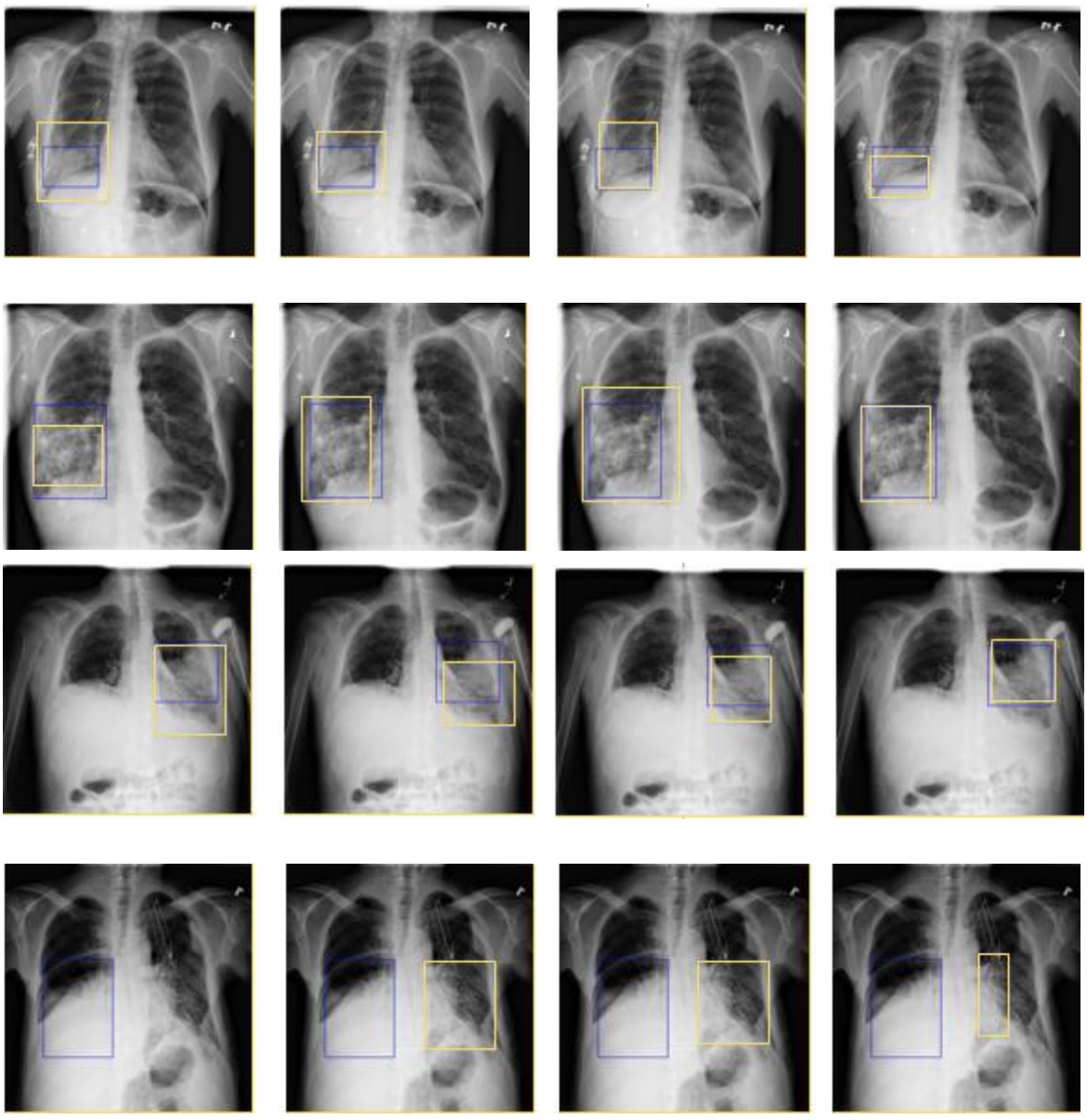
We controlled the learning intensity of the negative samples by setting different values of λ. When the value of λ is larger, the loss value of the negative samples is greater, and the algorithm's learning intensity of the negative samples is greater. As can be seen from Table 7, when $\lambda = 50$, the AP of the algorithm was still very low after 200k iterations because there were many negative samples being predicted as positive samples, resulting in low accuracy. When $\lambda = 200$, the algorithm had the highest accuracy after 90k iterations. When $\lambda = 250$, the prediction accuracy of the algorithm started to decrease after 90k iterations because the λ value was too large, and the gradient direction was basically dominated by negative samples, resulting in the poor learning of positive samples.

As Table 8 shows, the average precision of PYolo for different IOU thresholds was higher than other algorithms. Faster RCNN is a two-stage algorithm, while SSD, Yolov3 and PYolo are one-stage algorithms. Faster RCNN uses the RPN and it can control the proportion of positive and negative samples well; therefore, the average precision of Faster RCNN was higher than that of SSD and Yolov3. PYolo is an improvement on Yolov3 in feature fusion. Although its mean AP (mAP) was higher than Faster RCNN, PYolo was not able to avoid the imbalance of positive and negative samples. Figure 10 shows the detection effectiveness of the different algorithms. It can be seen that the localization

accuracy of PYolo and Yolov3 was higher than that of SSD and Yolov3, and that it had a slight advantage over Faster RCNN. However, for the last image, all four algorithms exhibited false detection and missed detection.

Table 8. Comparison of performance between PYolo and others (%).

Approach	AP@0.4	AP@0.5	AP@0.6	mAP
Faster RCNN [5]	58.63	43.76	26.64	43.01
SSD [29]	55.85	40.64	24.71	40.40
Yolov3 [8]	57.91	43.01	29.22	43.40
PYolo3	64.16	44.71	31.63	46.84



(**a**) (**b**) (**c**) (**d**)

Figure 10. Detection effectiveness of different algorithms. The blue box in each image is the ground-truth box, and the yellow box is the predicted box: (**a**) SSD; (**b**) Faster RCNN; (**c**) Yolov3; (**d**) PYolo.

The paired *t* test results show that PYolo has a significant improvement over Yolov3 in detection performance. Table 9 shows the test results for Yolov3 and PYolo for 600 images. During the testing

phase, the 600 images were divided into 10 groups and the detection accuracy of the algorithms was tested. The test statistic t was calculated to be 2.687, and $t(9)_{0.05}$ = 2.262 and $t(9)_{0.01}$ = 3.250 were determined by looking up tabulated t values. For |t| = 2.687, the p-value range was [0.01, 0.05]. According to this p-value range, the detection performance of PYolo was significantly better than that of Yolov3.

Table 9. Statistical significance (%).

Approach	mAP										Standard Deviation
Yolov3	42.7	44.8	35.0	42.8	45.8	47.3	42.0	45.6	47.0	41.1	3.64218
PYolo	46.6	52.0	37.2	43.9	48.6	47.0	45.9	49.3	51.8	46.0	4.04501

In order to compare the performance of PYolo with other pneumonia classification algorithms, the location information of each image detection result was ignored, and the accuracy of the classification results was determined. The criterion used was that if the confidence of at least one predicted bounding box was greater than or equal to 0.3, and the true label of the image was pneumonia, or the confidence of all predicted bounding boxes was less than 0.3, and the true label of the image was a normal lung, then the prediction result of the algorithms was judged to be correct. Table 10 shows the detection results for the different algorithms. The evaluation index was the ratio of the number of correctly classified images to the number of classified images. As Table 10 shows, CheXNet had the highest classification accuracy, because CheXNet uses the pre-training model provided in the study [13], but this model cannot provide location information. The classification accuracy of PYolo was 81.0, which was higher than that of the other algorithms. The feature extraction network of PYolo and Yolov3 is DarkNet53, which is deeper than VGG16 used by Faster RCNN, SSD and CNN + SVM.

Table 10. Classification accuracy of different algorithms (%).

Test	CNN + SVM [10]	CheXNet [13]	Faster RCNN [5]	SSD [29]	Yolov3 [8]	PYolo
Accuracy	70.8	83.7	79.5	75.0	77.5	81.0

4. Conclusions

This study firstly analyzes the problem of small characteristic differences in X-ray images of pneumonia lesions and proposes an improved end-to-end pneumonia detection algorithm based on Yolov3. The three main improvements offered by the proposed algorithm are as follows: the use of MaskFPN to generate pixel weights to suppress the output of inaccurate semantic information in low-level features, the use of a dilated convolution enhancement algorithm to detect pneumonia lesions, and the generation of a lesion anchor box with double K-means. The comparison experiment on the RSNA dataset proved that MaskFPN, dilated convolution and double K-means improved the detection ability of pneumonia lesions. We also demonstrated how to configure the parameters for MaskFPN and dilated convolution to facilitate further development of the algorithm.

Author Contributions: Conceptualization, S.Y. and Y.C.; methodology, S.Y. and X.T.; software, S.Y.; validation, S.Y., R.J. and S.M.; formal analysis, S.Y.; investigation, S.Y.; resources, S.Y.; data curation, S.M.; writing—original draft preparation, S.Y.; writing—review and editing, S.M.; visualization, S.Y.; supervision, X.T.; project administration, S.Y.; funding acquisition, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Funds for the Central Universities.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. U.S. Centers for Disease Control (CDC). Available online: https://www.cdc.gov/features/pneumonia/index.html (accessed on 7 December 2019).

2. Kaggle. Available online: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data (accessed on 10 November 2019).
3. Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
4. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 1440–1448.
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
6. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, R. You only look once: Unified, real-time object detection. *arXiv* **2015**, arXiv:1506.02640.
7. Redmon, J.; Farhadi, A. Yolo9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
8. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
10. Varshni, D.; Thakral, K.; Agarwal, L.; Nijhawan, R.; Mittal, A. Pneumonia detection using CNN based feature extraction. In Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 20–22 February 2019; pp. 1–7.
11. Setio, A.A.A. Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE Trans. Med. Imaging* **2016**, *35*, 1160–1169. [CrossRef] [PubMed]
12. Armato, S.G. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef] [PubMed]
13. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* **2017**, arXiv:1711.05225.
14. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
15. Chen, L.-C.; Papandreou, G.; Schroff, F.; Hartwing, A. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
16. Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the CVPR2017, Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
17. Tian, C.; Xu, Y.; Li, X.; Zuo, W.; Fei, L.; Liu, H. Attention-guided CNN for image denoising. *Neural Netw.* **2020**, *124*, 117–129. [CrossRef] [PubMed]
18. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.S.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the ICML2015, Lille, France, 6–11 July 2015; pp. 2048–2057.
19. Long, J.; Shelhamer, E.; Darrel, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
20. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
21. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2999–3007.
22. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
23. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
24. Srivastava, R.K.; Gref, K.; Schmidhuber, J. Highway networks. *arXiv* **2015**, arXiv:1505.00387.

25. Wang, B.B.; Wang, Y.X. Some properties of stochastic gradient descent method. *J. Math.* **2011**, *31*, 1041–1044.
26. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural Netw.* **1999**, *12*, 145–151. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015; pp. 1026–1034.
28. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
29. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *arXiv* **2015**, arXiv:1512.02325.