SSD-VUPV: A Novel SSD-Based Model for Automated Pulmonary Nodule Detection

ZHANLIN JI[1], JIANHUA PANG[2], JIANUO LIU[2], CHENXU DAI[2], SHENGNAN HAO[2,*], IVAN GANCHEV[3,4,5,*]

[1]Key Laboratory of Intelligent Forestry Monitoring and Information Technology,
Zhejiang A&F University,
HangZhou 311300,
CHINA

[2]Department of Artificial Intelligence,
North China University of Science and Technology,
Tangshan 063009,
CHINA

[3]TRC/ECE, University of Limerick,
Limerick, V94 T9PX,
IRELAND

[4]Faculty of Mathematics and Informatics,
University of Plovdiv "Paisii Hilendarski",
Plovdiv 4000,
BULGARIA

[5]Institute of Mathematics and Informatics—Bulgarian Academy of Sciences (IMI-BAS),
Sofia 1040,
BULGARIA

**Corresponding authors*

Abstract: - Purpose: In the context of increasingly scarce medical resources, the main purpose of this paper is to propose a SSD-based Variance-based Upsampling and Pyramid Voting (SSD-VUPV) model for improving the efficiency and accuracy of automated detection of lung nodules, as one of the early manifestations of lung cancer. Methods: Firstly, the proposed SSD-VUPV model adapts to the detection of small pulmonary nodules by changing the size of the feature map of the input prediction module. Secondly, the prediction frame is modified to make greater use of the shallow feature layer. In addition, a set of up-Block modules is augmented through the incorporation of asymmetric convolutions, which involves the utilization of a Feature Pyramid Network (FPN) mechanism. Finally, multi-scale and asymmetric convolutions are added to the model to further improve its detection performance. Results: Experimental results, obtained on two public datasets, show that SSD-VUPV can increase the *mAP@0.5*, *F1 score*, and *sensitivity* of the baseline model (SSD) from 63.01% to 87.52%, 64.20% to 90.24%, and 63.25% to 88.74%, on the LUNA16 dataset, and from 77.6% to 83.2%, 76.5% to 83.0%, and 75.8% to 81.9%, on the ChestX-ray14 dataset, respectively. Moreover, SSD-VUPV outperforms state-of-the-art models, based on their results reported in the literature, according to all evaluation metrics used. Conclusion: By cleverly integrating a feature pyramid structure and incorporating the newly designed up_Block modules, the proposed SSD-VUPV model can combine deep semantic features with shallow detail features, thus fully leveraging the rich feature information in the medical images, which allows it to reach top detection accuracy and robustness. Moreover, the inclusion of the newly designed Visual Multi-scale Asymmetric Convolution (VMAC) modules enables the model to adapt to different scale receptive fields, which enables it to capture more varied and detailed features, deepening its understanding of the input data and significantly improving its ability to capture features of various sizes. Consequently, the proposed SSD-VUPV model exhibits improved performance in scenarios involving complex backgrounds and targets.

Key-Words: - asymmetric convolution; deep learning; detection of pulmonary nodules; feature pyramid; SSD; LUNA16; ChestX-ray14.

Received: September 23, 2025. Revised: December 17, 2025. Accepted: January 29, 2026. Published: April 3, 2026.

1 Introduction

In China, lung cancer ranks top among malignant tumors in terms of both incidence rate and mortality. According to [1], even though the five-year survival rate of patients suffering from lung cancer in China increases each year over the past 20 years, this rate still did not exceed 20% and is much less than the survival rates of patients suffering from other tumors. The same study indicates that the 5-year survival rate of lung cancer patients may decrease if the number of diagnostic stages increases. Typically, patients in 'stage I' have a survival rate of 55.5%, while those in 'stage IV' have a survival rate of only 5.3%. Given the possibility that some pulmonary nodules may turn into lung tumors and may deteriorate into lung cancer in later stages, using a simple testing method to identify pulmonary nodules quickly and accurately can provide patients with appropriate treatment solutions and enable them to have healthier bodies and more chances for survival.

Convolutional neural networks (CNNs) play a crucial role in the advancement of science and technology, demonstrating remarkable success in detecting pulmonary nodules in computed tomography (CT) images. Various deep learning models have been proposed for this purpose. Although few researchers have used the Single Shot MultiBox Detector (SSD) model [2] for lung nodule detection, our experiments, described further in this paper, show that SSD (with additional improvements) can perform very well in detecting lung nodules.

Considering practical situations, achieving an ideal state of providing complete three-dimensional (3D) lung images to a medical self-test system, operating on an Internet of Things (IoT) platform [3] for convenient use by patients and physicians (which is the ultimate goal of our research), is a challenging task. The parameter counts of 3D convolution kernels are much greater than those of two-dimensional (2D) convolutional kernels of the same size, resulting in an excessively large model size, which imposes excessively high hardware requirements on such a self-test system. Rather than relentlessly pursuing extreme detection performance at all costs, we believe that a more lightweight 2D convolution should be adopted for use in such a system, in order to reduce the substantial hardware upgrade requirements imposed by 3D convolution models on hospitals and medical laboratories. Therefore, we decided to employ traditional 2D image detection.

By reviewing the related work in this area, presented further in Subsection 2.2, we found the following issues in the existing methods for pulmonary nodule detection: (1) the processing approaches for lung CT image data vary widely among current methods, leading to notable differences in their outcomes; and (2) the majority of the detection methods do not fully utilize shallow-level feature information.

Addressing the aforementioned issues, this paper explores a comprehensive CT image transformation toolkit, encompassing a series of processes such as lung parenchyma extraction, coordinate transformation, pixel spacing adjustment, mean value normalization, etc., applied to medical images and their annotated files, which allows the creation of improved datasets. The focus of this paper is on leveraging information within shallow-level networks, including methods like increasing feature layer dimensions used for enhanced SSD pulmonary nodule detection, and integrating information from networks of varying depths, aiming to provide a higher-precision detection tool.

The main contributions of this paper can be summarized as follows:

- Newly designed up-sampling modules are added to SSD to improve its pulmonary nodule detection ability through feature fusion of deep- and shallow feature layers;
- A novel Visual Multi-scale Asymmetric Convolution (VMAC) module is proposed for inclusion into SSD to improve its feature extraction capability of the shallow feature layer by means of asymmetric convolution;
- A squeeze-and-excitation (SE) attention module is added to SSD to further enhance its detection performance;
- Targeted adjustments are made to the *default* frame size and aspect ratio of the SSD anchors in order to tailor these parameters to the characteristics of pulmonary nodules contained in medical images.

- By adding these novel components to SSD, a novel model, named SSD-based Variance-based Upsampling and Pyramid Voting (SSD-VUPV), is formed and proposed for automated pulmonary nodule detection, as described further in this paper.
- The rest of this paper is structured as follows. Section 2 introduces materials and methods used w.r.t. the relevant technical background, related work done in the areas of pulmonary nodule detection and object detection in general, and the proposed SSD-VUPV model, which is described in detail. Section 3 presents the experimental results obtained, along with their analysis. Finally, Section 4 presents the concluding discussion.

2 Materials and Methods

2.1 Background

2.1.1 Receptive field

With the continuous changes in deep neural network (DNN) architectures and the deepening the networks, the receptive field of deep features is becoming larger, and the semantic expression capability is becoming stronger. A shallow neural network has a relatively small receptive field. Therefore, although its semantic expression ability is weak, it retains more detailed features because it does not undergo convolution operations in the deep network to blur many detailed features. Given the fact that pulmonary nodules, contained in the images of the LUNA16 dataset (used in the experiments described in Section 3), are small objects, the model proposed in the current paper utilizes a shallow network, whereby the 1×1×512 feature layer of SSD is replaced by the 75×75×256 layer of VGG16.

2.1.2 Feature Pyramid Network

In [4], the authors proposed the Deconvolutional Single Shot Detector (DSSD) model in 2017. Compared with the original SSD model, DSSD introduced two improvements in the network architecture:

1. Up-sampling modules are added;
2. A corresponding prediction head is added to each feature layer.

With reference to the DSSD model, up-sampling modules, based on bilinear interpolation, are utilized by the model proposed in this paper, each of which is fused with the feature map of the corresponding scale in the down-sampling module to form an encoder–decoder network architecture, thus enabling the structure of a feature pyramid network (FPN), [5], [6]. This combines the strong semantic information representation in the deep feature map with the detailed feature expression in the shallow feature map, which allows for enriching the prediction regression box and feature maps of different scales for the classification task, thereby enhancing the detection accuracy.

Residual networks [7] were developed to solve the problem of gradient disappearing or gradient explosion when the number of model layers increases. The concept of the residual network is introduced to the VMAC module proposed in this paper to improve the feature extraction capability of the shallow feature layer. Although methods such as data standardization, loading pre-training weights, and batch normalization were used several times during the experiments conducted by us to decelerate this trend, residual networks should still be ultimately introduced into the VMAC module.

2.1.3 Asymmetric Convolution

In the SSD-VUPV model, proposed in this paper, one 1×1, three 3×3 (with different expansion coefficients), and one 5×5 VGG16 convolutional module, as well as one VGG16 convolutional module that handles the same scale, are combined into one module for multi-scale convolution.

Asymmetric convolution [8] is utilized to cope with the heavy computation workload resulting from the use of one 5×5 or several 3×3 rectangular convolutions. The authors of [9] found a low-rank approximation, based on SVD decomposition, and then refined the upper layer to restore performance. [10] successfully learned horizontal and vertical kernels by minimizing reconstruction errors. The authors of [11] made 2D convolutions separable by applying structural constraints, thus accelerating the duration by twice while achieving considerable precision. On the other hand, asymmetric convolution is also widely used in CNN network architecture design. For instance, in the third version of Inception [12], the 7×7 convolution is replaced by a 1×7 convolution and a 7×1 convolution. Asymmetric convolution is also used in the DNN architecture ENet [13] as a sequence of 5×1 and 1×5 convolutions, which allows to increase the variety of functions learned by blocks and increase the receptive field.

Convolutional layer with a kernel size of $H \times W$ and D filters, using a C-channel feature map as an input, , produces an output $O \in \mathbb{R}^{R \times T \times D}$ with D channels, whereby the corresponding output feature map channel for the j-th filter is [8]:

$$O_{:,:,j} = \sum_{k=1}^{C} M_{:,:,k} * F_{:,:,k}^{(j)}, \quad (1)$$

where * denotes the 2D convolution operator, $M_{:,:,k}$ denotes the k-th channel of input $M \in R\hat{}(U \times V \times C)$, and $F_{:,:,k}^{(j)}$ denotes the k-th input channel of $F^{(j)}$, where $F \in \mathbb{R}^{H \times W \times C}$ is the filter's 3D convolution kernel.

In recent CNN architectures, batch normalization (BN) is commonly employed to address overfitting issues and speed up the training process. Compared to (1), the output channels then become, [8]:

$$O_{:,:,j} = \left(\sum_{k=1}^{C} M_{:,:,k} * F_{:,:,k}^{(j)} - \mu_j\right)\frac{\gamma_j}{\sigma_j} + \beta_j, \quad (2)$$

where μ_j and σ_j denote respectively the channel-wise mean and standard deviation of BN, and γ_j and β_j denote the learned scaling factor and bias term, respectively.

Standard convolutions can be substituted with equivalent asymmetric convolutions without introducing additional inference computational load. This is a useful property of convolution, meaning that the additivity may hold for 2D convolutions, even with different kernel sizes, [8]:

$$I * K^{(1)} + I * K^{(2)} = I * (K^{(1)} \oplus K^{(2)}), \quad (3)$$

where I denotes a matrix, $K^{(1)}$ and $K^{(2)}$ denote two 2D kernels with compatible sizes, and $\oplus$ denotes the element-wise addition of the kernel parameters on the corresponding positions.

So, this implies that one can augment smaller convolution kernels on top of larger ones. Formally, this kind of transformation on layer p and q is feasible, as stated in [8], if:

$$M^{(p)} = M^{(q)}, H_p \leq H_q, W_p \leq W_q, D_p = D_q \ , \quad (4)$$

For instance, using convolution kernels of size 3×1 and 1×3 has the same effect as using a 3×3 convolution kernel.

This can be confirmed by examining the computational process of standard sliding window convolutions, as shown in [8]. For a certain filter with kernel $F^{(j)}$, a certain point y on the output channel $O_{:,:,j}$ is given by:

$$y = \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} F_{h,w,c}^{(j)} X_{h,w,c} \ , \quad (5)$$

where X denotes the corresponding sliding window on input M. Obviously, when two output channels produced by two filters are summed up, the additivity (3) holds if for each point y on one channel, its corresponding point on the other channel shares the same sliding window X, [8].

2.1.4 Attention Mechanisms

Attention mechanisms serve as an important addition to deep-learning-based data processing schemes. The SE attention module [14] predicts a constant weight for each output channel and then calculates the weight for each channel. It performs a global pooling operation on the input features, makes adjustments by using fully connected and activated functions, and ultimately multiplies the functions with the input feature map. SE focuses more on the relationships between different channels and expects to understand the importance of various features in different channels. The advantage of SE lies in that it can be easily integrated with different neural networks to improve their performance using less computational power.

2.2 Related Work

2.2.1 Pulmonary Nodule Detection

The introduction of Computer-Aided Detection (CAD) systems has provided strong support for radiologists, assisting them in analyzing medical images, saving time and costs, reducing subjective diagnostic errors, and enhancing the accuracy and efficiency of lung cancer screening. Lung nodule segmentation methods are typically categorized into two main types: traditional methods and deep learning methods. Traditional methods primarily focus on manual feature extraction, employing techniques such as morphological features [15], pixel thresholding [16], and region growing [17]. The features encompass various low-level attributes like color, texture, shape, and gradient, either individually or in combinations. Through the extraction and analysis of these intermediate-level features, traditional methods achieve segmentation of candidate lung-nodule regions. In contrast, deep learning methods have the ability to automatically extract features, enabling the better detection of lung nodules.

In [18], the authors designed a model based on the location, shape, size, and gradient features of pulmonary nodules. The authors of [19] designed a fully automatic lung parenchyma segmentation model based on watershed algorithm, and customized feature analysis of lung texture features using the Hessian matrix. [20] proposed corresponding algorithms based on 128 expression features extracted from context information by using Sequential Floating Forward Selection (SFFS) and linear discriminant classification selection.

In [21], the authors introduced an improved residual network structure based on Faster R-CNN [22], which integrates depth-wise separable

convolution and pre-activation techniques, significantly enhancing the *sensitivity* of lung nodule detection. The authors of [23] made structural adjustments to Faster R-CNN, incorporating a deconvolutional layer and two-region proposal networks for detecting nodule candidates. However, despite achieving notable improvements, there were still cases of smaller nodules being missed. In [24], the authors employed a CNN for lung nodule detection and proposed a detection system based on deformable convolutions. A key aspect of their work was the addition of a branch network, trained to learn offsets from objects. Table 1 provides a summary of the presented related works.

Table 1. A summary of related works reviewed

Literature source	Based on	Key technical points
[21]	Faster R-CNN	Integrates depth-wise separable convolution and pre-activation techniques
[23]	Faster R-CNN	Incorporates a deconvolutional layer and two region proposal networks
[24]	SSD	Uses deformable convolutions
[18]	Fisher Linear Discriminant	Focuses on location, shape, size, and gradient features
[19]	MeVisPULMO	Watershed algorithm
[20]	Subsolid CAD system	Sequential Floating Forward Selection (SFFS) and linear discriminant classification selection

2.2.2 SSD

The SSD model appeared in a paper [2], published in 2016. For images of size 300×300 pixels each, SSD can achieve a *mAP@0.5* value of 74.3% on the PASCAL Visual Object Classes (VOC) 2007 dataset. In the case of the input images' size of 512×512 pixels, the *mAP@0.5* of 73.2% of the strongest model at that time (Faster R-CNN [22]) was surpassed by SSD, achieving a *mAP@0.5* value of 76.9%, [2].

As can be observed from the SSD network structure depicted in Figure 1 (Appendix), all parts in front of the third convolutional layer of Conv5 in VGG16 are used in the SSD's backbone network. In the example presented in Figure 1 (Appendix), the 300×300-pixel images are input and processed through three channels, followed by feature extraction performed through VGG16. The feature map of the Conv4 computation result in VGG16 (with a scale of 38×38) is taken out and passed through six feature layers (38×38, 19×19, 10×10, 5×5, 3×3, 1×1). Among these, the 38×38 feature layer is used for detecting relatively small objects, while the 1×1 feature layer is used to detect large objects.

The *default* values of the frame size and aspect ratio used by each feature layer of SSD are listed in Table 2. There are six groups of scales shown in the first row, whereby each feature layer corresponds to the preset scale S_k for the prediction of the feature layer and the preset scale S_{k+1} for the prediction of the next feature layer. For example, the preset scale of the second feature layer is $S_k = 45$, and the preset scale of the third feature layer is $S_{k+1} = 99$. The second row, which consists of six groups of tuples, shows the preset aspect ratio for each feature layer. For example, the aspect for the second feature layer is 1, 2, 0.5, 3, and 0.33. It should be noted that for a feature map with a preset scale of 45, an option with the side length of $\sqrt{45 \times 99}$ needs to be added with an aspect of 1. Thus, for the 19×19 feature map, there are a total of six different *default* boxes in the feature map, as detailed in Table 3.

Table 2. The default values of the SSD parameters (short list)

SSD parameter	*Default* value
Frame size	[(21,45),(45,99),(99,153),(153,207),(207,261),(261,315)]
Aspect ratio	[(1,2,0.5), (1,2,0.5,3,0.33), (1,2,0.5,3,0.33), (1,2,0.5,3,0.33), (1,2,0.5), (1,2,0.5)]

Table 3. The default values of the SSD parameters (extended list)

Feature Map No.	Shape	Size	Quantity
1	38×38	$21\{\frac{1}{2},1,2\};\sqrt{21\times45}\{1\}$	38×38×4=5776
2	19×19	$45\{\frac{1}{3},\frac{1}{2},1,2,3\};\sqrt{45\times99}\{1\}$	19×19×6=2166
3	10×10	$99\{\frac{1}{3},\frac{1}{2},1,2,3\};\sqrt{99\times153}\{1\}$	10×10×6=600
4	5×5	$153\{\frac{1}{3},\frac{1}{2},1,2,3\};\sqrt{153\times207}\{1\}$	5×5×6=150
5	3×3	$207\{\frac{1}{2},1,2\};\sqrt{207\times261}\{1\}$	3×3×4=36
6	1×1	$261\{\frac{1}{2},1,2\};\sqrt{261\times315}\{1\}$	1×1×4=4

According to data in Table 3, the total number of *default* boxes for the six prediction feature layers of SSD is equal to 8732 (5776+2166+600+150+36+4).

2.3 Proposed Model: SSD-VUPV

2.3.1 Model Overview

In the original SSD model, after the feature maps are downsampled to a set of specific sizes, they are directly fed into classifiers and regressors for target prediction. While this processing flow is simple and efficient, it overlooks the rich hierarchical information and detailed features in shallow layers, limiting the model's performance in handling complex scenes. To overcome this limitation and fully utilize multi-scale information in images, inspired by the DSSD model [4] and U-Net model [25] network structures, we propose introducing upsampling modules, named up_Block modules, into the SSD model. This improved design aims to incorporate an FPN structure under the multi-scale concept into the SSD model. By incorporating up_Block modules, the model can fuse information from different depth layers so as to leverage deep- and shallow feature information more effectively.

Deep feature maps contain rich semantic information, while shallow feature maps retain more pixel-level details. The combination of these two types of information provides the model with more comprehensive data interpretation capabilities. The overall structure of the proposed SSD-VUPV model is shown in Figure 2 (Appendix).

The SSD-VUPV model also utilizes newly designed Visual Multiscale Asymmetric Convolution (VMAC) modules, which are based on the concept of parallel multi-branch structures and asymmetric convolutions under the multi-scale idea. VMAC enhances feature extraction capabilities during down-sampling. Moreover, the proposed model incorporates the SE attention mechanism to better capture relevant target features, reduce the interference of irrelevant information, and improve lung nodule detection. Finally, the size of the feature layer used for classification and regression is increased to make the model better adapt to the characteristics of datasets.

2.3.2 Improved SSD Feature Sizes

By analyzing the target annotation after data preprocessing, described further in Sub-section 3.2, it can be observed that the target size in the LUNA16 dataset, used in the experiments, accounts for a very small proportion of the image size, and the pulmonary nodules in the LIDC-IDRI 3D annotation are shaped like an eccentric ellipsoid, so that the image on a single layer of 1cm-thick slice tends to be a square. Although the SSD model is more suitable for the detection of small-sized targets than YOLO [26] and Faster R-CNN [22], its *default* frame size and aspect ratio are not quite suitable for detecting pulmonary nodules contained in the CT images of the LUNA16 dataset. Thus, the *default* values of these SSD parameters must be *adjusted* to better suit this dataset used in the experiments conducted. Adjustments details are shown in Table 4.

Table 4. The *adjusted* values of the SSD parameters (short list), used by the proposed SSD-VUPV model

SSD parameter	***Adjusted* value**
Frame size	[(4,4), (8,8), (16,16), (20,20), (24,24), (30,30), (36,36)]
Aspect ratio	[(1), (1), (1), (1), (1), (1)]
Feature quantity	75×75,38×38,19×19,10×10,5×5,3×3

According to Table 2 and Table 3, the feature layer used for prediction in the original SSD model has a maximum size of 38×38 and a minimum size of 1×1, whereby the former is used for detecting small objects, and the latter is used for detecting large objects. Due to the existence of lots of small objects in the LUNA16 dataset as a whole, the 75×75 feature map of VGG16 was added and the original SSD's 1×1 feature map was removed, resulting in the proposed SSD-VUPV model (c.f., Figure 2 in Appendix), which enables increasing the total number of *default* boxes from 8,732 to 33,178 and improving the model detection of small objects, i.e., multidimensional pulmonary nodules in CT images.

2.3.3 Introducing up_Block modules

The small-scale feature map is un-pooled and upsampled, based on bilinear interpolation, and expanded to a size similar to that of the feature map of the previous layer, and then passed through a convolutional layer. The convolutional layer is primarily used to accurately control the size of the feature map obtained through unpooling to match the size of the feature map at the previous layer. Also, this layer is used to control the number of channels in the feature map, obtained through up-pooling, to ensure it matches the number of channels in the feature map of the previous layer, in order to facilitate the subsequent superposition of the two. For instance, a 10×10 feature map is changed to a 20×20 feature map through un-pooling and resized to a 19×19 feature map in which the convolution kernel step size is 2, the stride is 1, and the padding is 0. Meanwhile, the number of channels is adjusted to 1024. If the size of the feature map after up-sampling is consistent with the target size, convolution can only be used for

adjusting the number of channels. The up_Block module is shown in Figure 3 (Appendix).

2.3.4 Introducing VMAC modules

Based on the characteristics of multi-scale asymmetric convolution kernels, it can be deduced that there is more target information in the shallow feature layer, [27]. Based on this, a novel VMAC module was designed for use in the shallow feature layer. A combination of traditional standard convolution and asymmetric convolution is adopted in the VMAC module, as illustrated in Figure 4 (Appendix).

For the VMAC module, the VGG16-convx layer represents a convolution module of one of the conv1-5 layers. If the input image is of the original 300×300-pixel size, the two max-pooling operations are skipped, and the image directly enters the two 3×3 convolution kernels, shown on the left-hand and right-hand side of Figure 4 (Appendix). On the right-hand side, the results of five asymmetric convolution groups (at multiple scales) and those of the standard convolution are superposed, with an attempt to retain as many features as possible from each convolution operation. Then, the number of channels is sorted out through a 1×1 convolution kernel. The concept of residual linking is utilized for a partial residual operation in the VMAC module. Finally, the feature extraction results of the VGG16-convx layer are added and output through the SE attention layer.

3 Results and Analysis

3.1 Datasets

Two publicly available datasets were used in the experiments: – (1) the Lung Nodule Analysis 2016 (LUNA16) dataset – for model performance comparison; and (2) the ChestX-ray14 dataset – for studying the model generalizability.

LUNA16 [28], [29] is a subset of the largest common pulmonary nodule LIDC-IDRI dataset, [30]. LIDC-IDRI includes a total of 1,018 low-dose lung CT images and corresponding annotation information in *.xml* format, with unevenly distributed data samples, though. After deleting the CT images containing pulmonary nodules with a slice thickness larger than 3 mm and a maximum diameter less than 3 mm, a total of 888 low-dose lung CT images remained, which were stored in *.mhd* and *.raw* files. For the 5765 nodule regions in LUNA16, if two pulmonary nodules are found to be too close to each other (which is defined as having a distance of less than the sum of the radii or they are intersecting), they are merged, whereby their combined center and radius would be the average of the two. This way, a total of 2290 nodules remained. These nodules were jointly labeled by a total of four experts, whereby 2290, 1602, 1186, and 777 nodules were labeled by at least 1, 2, 3, or 4 experts, respectively. In the experiments, only the 1186 nodules, labeled by at least three experts, were used. Since CT images cannot directly enter a CNN for model training, data preprocessing was performed on this dataset, as described in the next subsection.

ChestX-ray14 [31] is a medical imaging dataset, comprising 112,120 frontal-view X-ray images of 30,805 patients with 14 common disease labels, mined from the text radiological reports via NLP techniques. It expands on the ChestX-ray8 dataset [32] by adding six additional thorax diseases, namely Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia.

3.2 Data Preprocessing on LUNA16

3.2.1 Hounsfield Unit Range

The Hounsfield unit (HU) value of a CT image is a measure of the density of a local tissue or organ in the human body corresponding to the extent to which the tissue/organ absorbs X-rays. The degree of absorption of X-rays by water is usually taken as a reference (i.e., the HU value of water is equal to 0), whereas the HU values of air (-1000) and of dense bone $(+1000)$ are used as the lower and upper boundary, respectively. The HU value of a CT lung image at the end of deep inhalation is equal to (-886 ± 22) HU, and at the end of deep exhalation it is equal to (-699 ± 52). In the experiments, HU values between $[-1000, 400]$ were retained for further data preprocessing as these are the thresholds commonly used in the LUNA16 competitions for normalization (any other object in a CT image with a HU value greater than 400 is usually just a bone).

3.2.2 Making Labels

Next, it was necessary to convert the pulmonary nodules' positions in the CT images into world coordinates and voxel coordinates, using the following formula:

$$nodule's\ position\ (in\ CT\ image) = nodule's\ position\ (in\ world\ coordinates) - the\ origin\ of\ the\ image\ coordinates$$

$$\begin{bmatrix} N_{(i,j)_x} \\ N_{(i,j)_y} \\ N_{(i,j)_z} \end{bmatrix} = \begin{bmatrix} W_{(i,j)_x} \\ W_{(i,j)_y} \\ W_{(i,j)_z} \end{bmatrix} - \begin{bmatrix} O_x \\ O_y \\ O_z \end{bmatrix}, \quad (6)$$

where $N_{(i,j)}$ denotes a pixel in the nodule's position (in a CT image), $W_{(i,j)}$ denotes a pixel in the nodule's position (in world coordinates), $O_{(i,j)}$ denotes a pixel in the origin of the image coordinates, and x, y, and z denote the coordinate values of the three dimensions of the pixel / origin.

Then, it was also necessary to correct the direction of the XYZ axes of the CT images since some images were taken when the patient had been lying down, and some – when s/he had been standing. For this, first, the corresponding *.xml* and *.mhd* files were found based on the ID number of the image. Then, a 3D contour mask map of the nodule was drawn based on the 3D coordinates and nodule diameter identification in the *.xml* file. The mask was converted into the relevant label format required by the corresponding model, after which the mask map of the pulmonary nodule was finally obtained.

3.2.3 Resampling

Further on, each CT image was multiplied with the corresponding lung mask to obtain the lung parenchyma area, whereby impurities were removed. Noticeably, the lung mask provided by LUNA16 was divided into left and right lungs. The actual length of the pixel spacing in the left lung mask image is 3 cm, while the actual length in the right lung mask image is 4 cm. To ensure consistent pixel spacing for the mask images of the left and right lungs, and facilitate subsequent feature extraction operations, we adjusted the pixel spacing of the masks to 1 cm; the value of the area outside the mask was set to zero for the subsequent model training.

3.2.4 Retaining Lung Parenchyma

As the pixel spacing in different CT images obtained differs with the unit of spacing not equal to 1, it was necessary to adjust the pixel spacing of CT images to 1, based on existing labeling information, and set the slice thickness also to 1 cm. This allowed us to uniform all image pixel specifications obtained.

3.2.5 Normalization and Mean Value Removal

As the size/roughness of image pixels varies, due to scanning for example, the distance between slices may be different. As this may affect CNN performance, one needs to perform resampling to the same resolution to eliminate the difference in scanner resolution. Finally, the resampling result should be normalized according to the range of HU values. If the normalized value of a pixel at a specific point in an image exceeds 1, then the value of that pixel is set to 1. Otherwise, if the normalized value of a pixel is less than 0, then the pixel value is set to 0. The overall values of the normalized images in the LUNA16 dataset are relatively large. In simple terms, it means that all pixel values are subtracted by the mean value, which for the LUNA16 dataset is equal to 0.25. This value was subtracted from the normalized value and saved in *.png* format as the final processed data for model training.

The entire preprocessing, performed on the LUNA16 dataset, is depicted in Figure 5 (Appendix).

3.2.6 Statistical Analysis of Images

A visualization of the distribution of the proportion of pulmonary nodule area to the entire image area in the LUNA16 dataset is presented in Figure 6 (Appendix). The vertical axis represents the ratio of the nodule's height to that of the entire image, and the horizontal axis represents the ratio of the nodule's width to that of the entire image. From the distribution chart, it can be observed that the overall size of the lung nodule targets is relatively small.

Figure 6 (Appendix) visualizes the scale of the targets, but it does not effectively reflect the detailed proportions of targets of diverse sizes. After performing statistical analysis on the images processed from the LUNA16 dataset, a distribution chart of the ratio of lung nodule area to the total image area was obtained. As shown in Figure 7 (Appendix), instances with a ratio below 0.002 account for 67% of the entire dataset. In about 90% of the lung nodule samples, the ratio of the nodule area to the total image area is below 0.004. In the MS COCO dataset, targets of different sizes are defined as follows: those smaller than 32×32 pixels are marked as small targets, those between 32×32 and 96×96 pixels are marked as medium targets, and those larger than 96×96 pixels are marked as large targets. This provides statistical support for adjusting the *default* value of the parameter 'frame size' in the SSD model.

3.3 Evaluation Metrics

In the conducted experiments, the performance of the elaborated SSD-based models for detecting multidimensional pulmonary nodules in medical images was compared to that of the original SSD model, based on the mean average precision (mAP), which is widely used, c.f., [33], [34], to evaluate the comprehensive performance of a model in performing an object detection task, as this metric

represents the area enclosed by *precision* and *sensitivity*. Basically, a higher *mAP* value pursues a better balance between *precision* and *sensitivity*. The mAP value is calculated as follows:

$$mAP = \frac{\sum AP}{N_{class}} , \quad (7)$$

where N_{class} denotes the number of all classes of objects for detection and *AP* denotes the average precision, calculated as the area enclosed by the *precision (p)–sensitivity (s)* curve and the X axis as follows:

$$AP = \int_0^1 p(s) d_s , \quad (8)$$

Precision refers to the proportion of true positive samples in the detection results, calculated as follows:

$$precision = \frac{TP}{TP+FP} , \quad (9)$$

where *TP* denotes the number of medical images containing correctly detected pulmonary nodules (of a correctly identified class) with an Intersection over Union (*IoU*) ratio (between the detected box and the ground-truth box) larger than a certain threshold (set to 0.5 in the conducted experiments) and a confidence level greater than a certain threshold (set to 0.5 in the experiments), and *FP* denotes the number of medical images containing negative samples, which are incorrectly detected as positive samples.

Sensitivity refers to the correctly detected samples among all positive samples, calculated as follows:

$$sensitivity = \frac{TP}{TP+FN} , \quad (10)$$

where *FN* denotes the number of medical images containing positive samples of pulmonary nodules, which are incorrectly identified as negative samples by a model.

In the conducted experiments, *IoU* was set to 0.5, and *AP* was calculated for each class across all images, and then averaged for all classes, resulting in *mAP@0.5*. Another metric, used in the experiments, was *mAP@0.5:0.95*, which represents the average *mAP* at different *IoU* thresholds (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95).
The third metric used was the *F1 score*, which is the harmonic mean of *precision* and *sensitivity*, calculated as follows:

$$F1\ score = 2 \cdot \frac{precision \cdot \text{sensitivity}}{precision + \text{sensitivity}} , \quad (11)$$

3.4 Main Findings

3.4.1 Choice of Baseline

Firstly, we conducted experiments with Faster R-CNN [22], YOLOv7 [35], Deformable DETR [36], SSD [2], EfficientDet [37], CornerNet [38], and YOLOX-s [39] models, based on half of the LUNA16 dataset (i.e., using the CT images contained in the subset0-subset4 fold-er), in order to identify the prime candidates among these best-performing models for participation as a baseline in the performance comparison with the SSD-VUPV model, proposed in this paper. The results, obtained in these experiments using various open-source implementations, are shown in Table 5. For benchmark model testing, the weights trained on the same training set, VOC2007, were used. These pre-trained weight files come partly from GitHub and partly from cloud storage. We applied a uniform approach to all models. We compared the results while ensuring replicated others' network models to the maximum extent possible. Based on these experiments, SSD was chosen as a baseline due to its superiority to other models, as evident from Table 5.

Table 5. Pulmonary nodule detection performance of state-of-the-art models considered as baseline candidates

Model	*mAP@0.5* (%)	*F1 score* (%)	*Sensitivity* (%)
Faster R-CNN	53.00	49.30	50.88
YOLOv7	56.30	50.96	51.32
Deformable DETR	62.80	60.37	59.68
EfficientDet	62.20	62.90	61.90
CornerNet	58.10	63.50	61.40
YOLOX-s	62.21	62.88	61.30
SSD (*baseline*)	**63.01**	**64.20**	**63.25**

3.4.2 Adjusting SSD Parameter Values

Table 6 shows the results of adjusting the anchor frame size and aspect ratio in the proposed SSD-VUPV model versus the *default* frame size and aspect ratio of SSD. It can be seen that a good improvement in detection performance is achieved by each *adjusted* version compared to the corresponding *default* version, according to all evaluation metrics. The improvement effect becomes more obvious as the model gets more sophisticated. Therefore, the *adjusted* parameter values were used in the proposed model.

Table 6. SSD pulmonary nodule detection performance, averaged over five separately conducted experiments, using the SSD *default* and *adjusted* parameter values

Model	***mAP@0.5*** (%)	***F1 score*** (%)	***Sensitivity*** (%)
SSD (*default*)	63.01	64.20	63.25
SSD (*adjusted*)	64.76	65.83	65.29
SSD (*default*) + up_Block	75.32	78.91	77.35
SSD (*adjusted*) + up_Block	79.62	81.33	80.60
SSD (*default*) + up_Block + SE	80.08	81.84	80.32
SSD (*adjusted*) + up_Block + SE	**86.04**	**88.35**	**87.54**

3.4.3 Ablation Study

Table 7 presents detailed ablation study experiments conducted by using SSD (*default*) as a baseline. First, incorporating the SE attention module into the baseline improved *mAP@0.5*, *F1 score*, and *sensitivity* values by 2.22, 2.89, and 2.03 percentage points, respectively. By utilizing another type of attention module, CBAM, instead of SE, all metrics showed greater improvement by 3.93, 4.31, and 4.91 percentage points, respectively, compared to the baseline. However, as the network deepens with the addition of up-Block modules, we noticed that the effect of using CBAMs has started to decline, compared to that of using SEs. Therefore, we decided to abandon CBAMs and use only SE attention modules in the proposed model. Next, we evaluated the effect of using the upsampling module (up_Block) in combination with other components. Results, shown in Table 7, indicate that adding the up_Block module to SSD leads to significant improvements on all metrics, with the improvement magnitude stabilizing as network depth increases. Furthermore, incorporating the SE attention module allowed further improvement of the detection performance, reaching the highest values of 87.52%, 90.24%, and 88.74%, respectively for *mAP@0.5*, *F1 score*, and *sensitivity* with the 'SSD + up_Block + VMAC + SE' combination, resulting in the proposed SSD-VUPV model. Even though this was done at the expense of increasing the number of model parameters (from 23.25 million in SSD to 263.74 million in SSD-VUPV) and the model size (from 92.99 MB in SSD to 739.47 MB in SSD-VUPV), this seems acceptable considering the significant improvement in detection performance.

Table 7. Results of the ablation study experiments, conducted with SSD as a baseline

Model components	***mAP@0.5*** (%)	***F1 score*** (%)	***Sensitivity*** (%)
SSD (*baseline*)	63.01	64.20	63.25
SSD + SE	65.23	67.09	65.28
SSD + CBAM	66.94	68.51	68.16
SSD + up_Block	79.62	81.33	80.60
SSD + up_Block + SE	86.04	88.35	87.54
SSD + up_Block + CBAM	83.66	85.92	84.57
SSD + up_Block + VMAC	82.17	83.86	82.99
SSD + up_Block + VMAC + SE (i.e., *proposed* SSD-VUPV)	**87.52**	**90.24**	**88.74**
SSD + up_Block + VMAC + CBAM	84.75	86.71	86.00

3.4.4 Detection of Small- and Medium-size Pulmonary Nodules

The statistics of *mAP@0.5:0.95* detection performance of the proposed SSD-VUPV model w.r.t. small- and medium-size pulmonary nodules in the LUNA16 dataset is presented in Table 8. As one can observe from this table, SSD-VUPV performs much better that the baseline in detecting small- and medium-size targets, increasing the *mAP@0.5:0.95* value by 9.50 percentage points w.r.t. small-size pulmonary nodules, and by 21.90 percentage points w.r.t. medium-size pulmonary nodules. Visual comparison of pulmonary nodule detection abilities of the proposed model and the baseline, based on the LUNA16 dataset, is presented in Figure 8 (Appendix). Overall, SSD-VUPV not only significantly outperforms the baseline in overall detection ability but also makes notable progress in addressing missed detections and enhancing the detection ability w.r.t. small-size lung nodules, as can be seen in the second row of compared images demonstrating that SSD-VUPV is able to detect small-size lung nodules, missed by SSD, thereby exhibiting better performance w.r.t. false negative (FN) cases.

Table 8. Small- and medium-sized pulmonary nodule detection performance comparison with the baseline, based on the LUNA16 dataset

Model	***mAP@0.5:0.95*** (%) for detecting small-sized pulmonary nodules	mAP@0.5:0.95 (%) for detecting medium-sized pulmonary nodules
SSD (*baseline*)	22.80	41.30
SSD + up_Block	28.30	50.40
SSD-VUPV (*proposed*)	**32.30**	**63.20**

3.4.5 Detection Performance Comparison

Finally, we compared the proposed SSD-VUPV model with state-of-the-art models, including pipeline [33], YOLOv5 refined [34], YOLOv3 improved [40], Method1+pre [41], and the 3D convolutional model DR-ATTO(3D) [42], using their results reported in the literature for the LUNA16 dataset. As can be seen from Table 9, the proposed SSD-VUPV model outperforms all these state-of-the-art models, according to all evaluation metrics used. More specifically, it overpassed the first runner-up by 4.22 percentage points based on *mAP@0.5*, by 13.84 percentage points according to *F1 score*, and by 14.44 percentage points based on *sensitivity*.

Table 9. Pulmonary nodule detection performance comparison with state-of-the-art models, based on the LUNA16 dataset

Model	*mAP@0.5* (%)	*F1 score* (%)	*Sensitivity* (%)
pipeline [33]	63.70	-	74.30
YOLOv5 refined [34]	75.00	-	-
YOLOv3 improved [40]	73.90	-	-
Method1+pre [41]	83.30	-	-
DR-ATTO(3D) [42]	-	76.40	73.44
SSD-VUPV (*proposed*)	**87.52**	**90.24**	**88.74**

3.4.6 Generalizability Study

To study the generalizability of the proposed model, we conducted experiments on another dataset (ChestX-ray14), by comparing the SSD-VUPV with state-of-the-art models. As can be seen from Table 10, the proposed SSD-VUPV model outperforms all these models, according to all evaluation metrics used. More specifically, it overpassed the second-placed model by 5.6 percentage points based on *mAP@0.5*, by 3.0 percentage points according to *F1 score*, and by 1.7 percentage points based on *sensitivity*.

Table 10. Pulmonary nodule detection performance comparison with state-of-the-art models, based on the ChestX-ray14 dataset

Model	*mAP@0.5* (%)	*F1 score* (%)	*Sensitivity* (%)
YOLOv7	75.3	74.2	73.1
CornerNet	77.0	77.5	76.8
Mask R-CNN	72.6	71.1	70.9
Faster R-CNN	69.2	68.0	65.8
EfficientDet	76.9	77.5	76.1
SSD	77.6	76.5	75.8
SSD-VUPV (*proposed*)	**83.2**	**83.0**	**81.9**

In particular, the visualized results, shown in the third row of Figure 9 (Appendix), demonstrate that the baseline (SSD) is prone to false positive (FP) detection, and thus it may incorrectly identify normal lung tissue as pulmonary nodules. On the contrary, the proposed SSD-VUPV model avoids most of these false detections, thereby exhibiting better performance w.r.t. FP cases.

4 Discussion

One of the major challenges of using the LUNA16 dataset is the diversity and complexity of lung nodules. Lung nodules exhibit different shapes, sizes, and densities in CT images, with many nodules being very small and difficult to distinguish. This diversity makes accurate detection of lung nodules a challenging task, requiring models to effectively handle various types and sizes of nodules while maintaining high accuracy. Additionally, chest CT images typically contain a lot of anatomical structures and background noise, thus further increasing the difficulty of lung nodule detection and necessitating the use of models with strong robustness and resistance to interference.

The main advantage of the SSD model is that it can complete object detection in a single forward pass, making its inference speed very high and suitable for real-time applications. Additionally, it can use feature maps of different scales for detection, enhancing the detection of objects of various sizes. However, the SSD model has shortcomings in feature fusion. Although it uses feature maps of different scales, there is a lack of effective feature fusion between these feature maps. As a result, the deep semantic features obtained from a single forward pass cannot be effectively combined with the shallow primary features.

Feature pyramids enable models to effectively detect targets of varying sizes and enhance detection of small objects by allowing analysis of features at multiple scales. While the YOLO series incorporates the concept of feature pyramids, it does not fully apply it throughout the model design, merging only two feature map sizes and discarding some feature information. In contrast, the U-Net series fully embraces the idea of feature fusion. However, the original U-Net network architecture [25] is relatively shallow, typically consisting of only five layers of encoders and decoders, which limits its efficiency in feature extraction. Moreover, as U-Net focuses on pixel-level segmentation, it primarily uses a single, largest-size feature map for segmentation, which may restrict its ability to handle multi-scale features.

The SSD-VUPV model, proposed in this paper, cleverly integrates the feature pyramid structure through the incorporation of the newly designed up_Block modules, effectively utilizing six groups of feature layers of different scales for feature fusion. This design strategy allows the model to combine deep semantic features with shallow detail features, fully leveraging the rich feature information in images, thereby enhancing the model's performance and robustness. Moreover, equipped with nine encoder layers, the proposed model surpasses U-Net in terms of the structural depth, which further strengthens its ability to extract image features. The inclusion of the newly designed VMAC modules enables the model to adapt to different scale receptive fields. Parallel branches allow it to extract information through different paths within the same layer, with each branch utilizing different types of filters or convolution kernel sizes. This design enables the model to capture more varied and detailed features, deepening its understanding of the input data and significantly improving its ability to capture features of various sizes. Consequently, the SSD-VUPV model exhibits improved performance in scenarios involving complex backgrounds and targets. The experimental results, obtained on the LUNA16 and ChestX-ray14 datasets, demonstrate that SSD-VUPV outperforms state-of-the-art models in lung nodule detection.

Mainstream detection models generally exhibit low performance when processing completed lung images. This issue persists even with deeper models like YOLO [26] and Faster R-CNN [22]. Therefore, our future research plan includes enhancing the image quality by employing image denoising techniques such as speckle reduction, salt-and-pepper noise removal, and spatial denoising. Additionally, we aim to boost the performance of the proposed SSD-VUPV model by utilizing data augmentation techniques to perform various transformations and operations to the original images, thereby expanding the used datasets.

Another research direction involves model lightweighting, which includes techniques such as distillation and depth-wise separable convolution to reduce model size and increase its detection speed. The parameter count (263.74 M) and size (739.47 MB) of the proposed model are substantial compared to the baseline (23.25 M and 92.99 MB, respectively). Through logical analysis and code examination, we identified that the computation of the full channel count in the VMAC multi-scale convolution part was one of the main reasons for the excessive parameter count. If the channel count in this part could be reduced to one-fourth of the original count, the total parameter count would be decreased by 50%. However, reducing the number of channels in SSD-VUPV may result in a decrease of detection performance compared to the full-channel SSD-VUPV. In the future, we plan to conduct validation experiments and explore adequate lightweight strategies to tackle this issue. This approach can preserve hardware resources used by lung nodule detection systems, enhancing their ability to successfully serve doctors and patients.

References:

[1] He, J.; Li, N.; Chen, W.Q.; Wu, N.; Shen, H.B.; Jiang, Y.; Li, J.; Wang, F.; Tian, J.H. Consulting group of China guideline for the screening and early diagnosis and treatment of lung cancer, 43(3): 243-268. Beijing, 2021. doi: 10.3760/cma.j.cn112152-20210119-00060.

[2] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Lecture Notes in Computer Science, vol. 9905, pp. 21-37 *(14th European Conference on Computer Vision – ECCV 2016*, Amsterdam, The Netherlands, 11-14 October 2016). Springer, Cham. 2016. doi: 10.1007/978-3-319-46448-0_2.

[3] Ganchev, I.; Ji, Z.L.; O'Droma, M. Horizontal IoT Platform EMULSION. *Electronics*, 12(8):1864. 2023. doi: 10.3390/electronics12081864.

[4] Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional single shot detector. arXiv:1701.06659. 2017. doi: 10.48550/arXiv.1701.06659.

[5] Jensen, J.D.; Elewski, B.E. The ABCDEF rule: combining the “ABCDE rule” and the “ugly duckling sign” in an effort to improve patient self-screening examinations. *Journal of Clinical Aesthetic Dermatology*, 8(2): 15. 2015.

[6] Zhang, X.; Cao, S.; Chen, C. Scale-aware hierarchical detection network for pedestrian detection. *IEEE Access*, 8: 94429-94439. 2020. doi: 10.1109/ACCESS.2020.29953212020.

[7] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778. Las Vegas, NV, USA,

27-30 June 2016. doi: 10.1109/CVPR.2016.90.

[8] Ding, X.; Guo, Y.; Ding, G.; Han, J. ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks. *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1911-1920. Seoul, Korea (South), 27 Oct.-2 Nov. 2019. doi: 10.1109/ICCV.2019.00200.

[9] Denton, E.L.; Zaremba, W.; Bruna, J.; LeCun, Y.; Fergus, R. Exploiting linear structure within convolutional networks for efficient evaluation. *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'14)*, vol. 1, pp. 1269-1277. Montréal, Canada, 8-13 December 2014.

[10] Jaderberg, M.; Czarnecki, W.M.; Osindero, S.; Vinyals, O.; Graves, A.; Silver, D.; Kavukcuoglu, K. Decoupled neural interfaces using synthetic gradients. *Proceedings of the 34th International Conference on Machine Learning, pp. 1627-1635.* Sydney, Australia, 6-11 August 2017.

[11] Jin, J.; Dundar, A.; Culurciello, E. Flattened convolutional neural networks for feedforward acceleration. *Proc. of ICLR workshop*, San Diego, CA, USA. 2015.

[12] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818-2826. Las Vegas, NV, USA, 27-30 June 2016. doi: 10.1109/CVPR.2016.308.

[13] Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147. 2016. doi: 10.48550/arXiv.1606.02147.

[14] Zhang, H.; Zhao, W.; Liu, S. SE-ECGNet: A multi-scale deep residual network with squeeze-and-excitation module for ECG signal classification. *Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2685-2691. Seoul, Korea (South), 16-19 December 2020. doi: 10.1109/BIBM49941.2020.9313548.

[15] Okumura, T.; Miwa, T.; Kako, J.I.; Yamamoto, S.; Matsumoto, M.; Tateno, Y.; Iinuma, T.; Matsumoto, T. Automatic Detection of lung cancers in chest CT images by variable N-Quoit filter. *Proceedings of the 14th International Conference on Pattern Recognition*, vol. 2, pp. 1671-1673. Brisbane, QLD, Australia, 1998. doi: 10.1109/ICPR.1998.712041.

[16] Armato, S.G.; Giger, M.L.; Moran, C.J.; Blackburn, J.T.; Doi, K.; MacMahon, H. Computerized detection of pulmonary nodules on CT scans. *Radiographics*, 19(5): 1303-1311. 1999. doi: 10.1148/radiographics.19.5.g99se181303.

[17] Suiyuan, W.; Junfeng, W. Pulmonary nodules 3D detection on serial CT scans. *Proceedings of the 2012 3rd Global Congress on Intelligent Systems*, pp. 257-260. Wuhan, China, 6-8 November 2012. doi: 10.1109/GCIS.2012.46.

[18] Messay, T.; Hardie, R.C.; Rogers, S.K. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Med Image Anal*, 14(3): 390-406. 2010. doi: 10.1016/j.media.2010.02.004.

[19] Lassen, B.; Kuhnigk, J.-M.; Friman, O.; Krass, S.; Peitgen, H.-O. Automatic segmentation of lung lobes in CT images based on fissures, vessels, and bronchi. *Proceedings of the 2010 IEEE international symposium on biomedical imaging: from nano to macro*, pp. 560-563. Rotterdam, Netherlands, 14-17 April 2010. doi: 10.1109/ISBI.2010.5490284.

[20] Jacobs, C.; Van Rikxoort, E.M.; Twellmann, T.; Scholten, E.T.; De Jong, P.A.; Kuhnigk, J.M.; Oudkerk, M.; De Koning, H.J.; Prokop, M.; Schaefer-Prokop, C.; van Ginneken, B. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical Image Analysis*, 18(2): 374-384. 2013. doi: 10.1016/j.media.2013.12.001.

[21] Shi, L.; Ma, H.; Zhang, C.; Fan, S. Detection method of pulmonary nodules based on improved residual structure. *Journal of Computer Applications*, 2020(7): 2110-2116. 2020.

[22] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS'15)*, vol. 1, pp. 91-99. Montreal, Canada, 11-12 December 2015.

[23] Xie, H.; Yang, D.; Sun, N.; Chen, Z.; Zhang, Y. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognition*, 85: 109-119. 2019. doi: 10.1016/j.patcog.2018.07.031.

[24] Gu, J.; Tian, Z.; Qi, Y. Pulmonary nodules detection based on deformable convolution. *IEEE Access*, 8: 16302-16309. 2020. doi: 10.1109/ACCESS.2020.2967238.

[25] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds). *Lecture Notes in Computer Science*, vol 9351, pp. 234-241 (Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Munich, Germany, 5-9 October 2015). Springer, Cham. 2015. doi: 10.1007/978-3-319-24574-4_28.

[26] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: unified, real-time object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788. Las Vegas, NV, USA, 27-30 June 2016. doi: 10.1109/CVPR.2016.91.

[27] Dolz, J.; Desrosiers, C.; Ayed, I.B. IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet. *Proceedings of the 5th international workshop and challenge on computational methods and clinical applications for spine imaging (CSI 2018)*, pp. 130-143. Granada, Spain, 16 September 2018. doi: 10.1007/978-3-030-13736-6_11.

[28] Setio, A.A.; Jacobs, C.; Gelderblom, J.; van Ginneken, B. Automatic detection of large pulmonary solid nodules in thoracic CT images. *Medical Physics*, 42: 5642-5653. 2015. doi: 10.1118/1.4929562.

[29] Murphy, K.; van Ginneken, B.; Schilham, A.M.; de Hoop, B.J.; Gietema, H.A.; Prokop, M. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification. *Medical Image Analysis*, 13(5): 757-770. 2009. doi: 10.1016/j.media.2009.07.001.

[30] Armato III, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A.; Kazerooni, E.A.; MacMahon, H.; van Beek, E.J.R.; Yankelevitz, D.; Biancardi, A. M.; Bland, P.H.; Brown, M.S.; Engelmann, R.M.; Laderach, G.E.; Max, D.; Pais, R.C.; Qing, D.P.-Y.; Roberts, R.Y.; Smith, A.R.; Starkey, A.; Batra, P.; Caligiuri, P.; Farooqi, A.; Gladish, G.W.; Jude, C.M.; Munden, R.F.; Petkovska, I.; Quint, L.E.; Schwartz, L.H.; Sundaram, B.; Dodd, L.E.; Fenimore, C.; Gur, D.; Petrick, N.; Freymann, J.; Kirby, J.; Hughes, B.; Casteele, A.V.; Gupte, S.; Sallam, M.; Heath, M.D.; Kuhn, M. H.; Dharaiya, E.; Burns, R.; Fryd, D.S.; Salganicoff, M.; Anand, V.; Shreter, U.; Vastagh, S.; Croft, B.Y.; Clarke, L.P. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2): 915-931. 2011. doi: 10.1118/1.3528204.

[31] Alcázar, C. NIH Chest X-ray dataset, [Online]. https://paperswithcode.com/dataset/chestx-ray14 (Accessed Date: August 10, 2025)..

[32] Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462-3471. Honolulu, HI, USA, 21-26 July 2017. doi: 10.1109/CVPR.2017.369.

[33] Fang, D.; Jiang, H.; Chen, W.; Qin, Z.; Shi, J.; Zhang, J. Pulmonary nodule detection on lung parenchyma images using hyber-deep algorithm. *Heliyon*, 9(7): e17599. 2023. doi: 10.1016/j.heliyon.2023.e17599.

[34] Shao, Z.; Wang, G.; Zhou, C. Imageological Examination of Pulmonary Nodule Detection. *Proceedings of the 2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pp. 383-386. Zhuhai, China, 24-26 September 2021. doi: 10.1109/ICBASE53849.2021.00077.

[35] Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv:2207.02696, 2022. doi: 10.48550/arXiv.2207.02696.

[36] Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020. doi: 10.48550/arXiv.2010.04159.

[37] Tan, M.; Pang, R,; Le, Q.V. EfficientDet: scalable and efficient object detection. arXiv:1911.09070, 2020. doi: 10.48550/arXiv.1911.09070.

[38] Law, H.; Teng, Y.; Russakovsky, O.; Deng, J. CornerNet-Lite: efficient keypoint based

object detection. arXiv:1904.08900, 2019. doi: 10.48550/arXiv.1904.08900.

[39] Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: exceeding YOLOX series in 2021. arXiv:2107.08430, 2021. doi: 10.48550/arXiv.2107.08430.

[40] Haibo, L.; Shanli, T.; Shuang, S.; Haoran, L. An improved YOLOv3 algorithm for pulmonary nodule detection. *Proceedings of the 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, pp. 1068-1072. Shanghai, China, 18-20 June 2021. doi: 10.1109/IMCEC51613.2021.9482291.

[41] Zhou, Z.; Gou, F.; Tan, Y.; Wu, J. A cascaded multi-stage framework for automatic detection and segmentation of pulmonary nodules in developing countries. *IEEE Journal of Biomedical and Health Informatics*, 26(11): 5619-5630, 2022. doi: 10.1109/JBHI.2022.3198509.

[42] Qi, Z.; Tao, G.; Liu, F.; Gui, H.; Huang, W.C.; Zou, J.N. DR-ConvNeXt: a dilated residual network based on ConvNeXt for lung nodule features prediction. *Proceedings of the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2169-2174. Istanbul, Turkiye, 5-8 December 2023. doi: 10.1109/BIBM58861.2023.10385309.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Conceptualization, Z.J. and J.P.; methodology, J.P. and C.D.; validation, I.G.; formal analysis, J.P., C.D. and Z.J.; investigation, S.H. and C.D.; resources, S.H. and J.L., visualization, I.G. and J.L.; data curation, C.D. and J.L.; writing—original draft, J.P.; writing—review and editing, J.P. and I.G.; supervision, S.H. and Z.J.; project administration, Z.J. and I.G; funding acquisition, Z.J. and I.G. All authors have read and agreed to the published version of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This publication has emanated from joint research conducted with the financial support of the National Key Research and Development Program of China under Grant No. 2017YFE0135700 and the Bulgarian National Science Fund (BNSF) under Grant No. КП-06-ИП-КИТАЙ/1 (KP-06-IP-CHINA/1).

Ethics Approval

This paper does not contain any experiments with human participants or animals performed by any of the authors.

Data Deposition Information: This study is based on publicly available data from the LUNA16 dataset (https://luna16.grand-challenge.org/Data/) and the ChestX-ray14 dataset, [31].

Conflict of Interest

The authors declare no competing interests.

APPENDIX

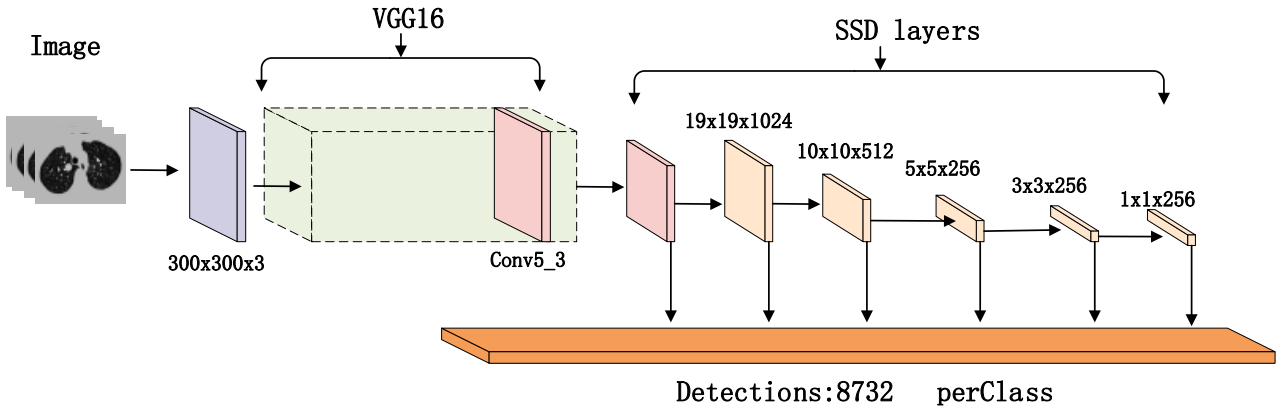


Fig. 1: The SSD network structure uses VGG16 as a backbone network
Source: created by the authors

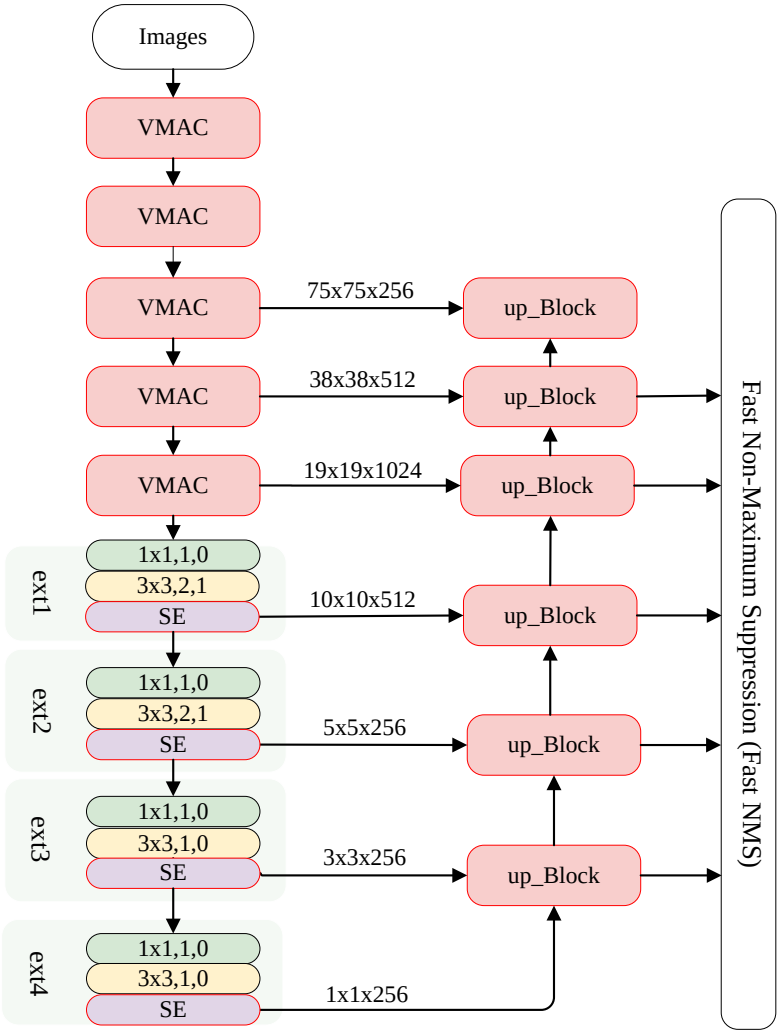


Fig. 2: The proposed SSD-VUPV model
Source: created by the authors

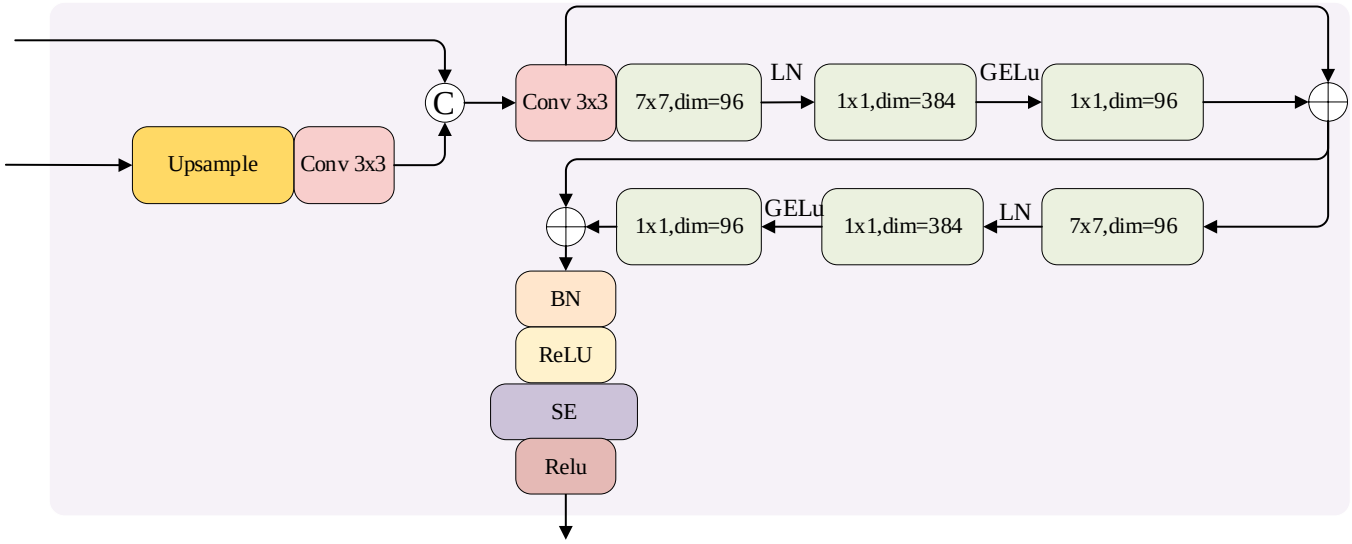


Fig. 3: The newly designed up_Block module, utilized by the proposed SSD-VUPV model
Source: created by the authors

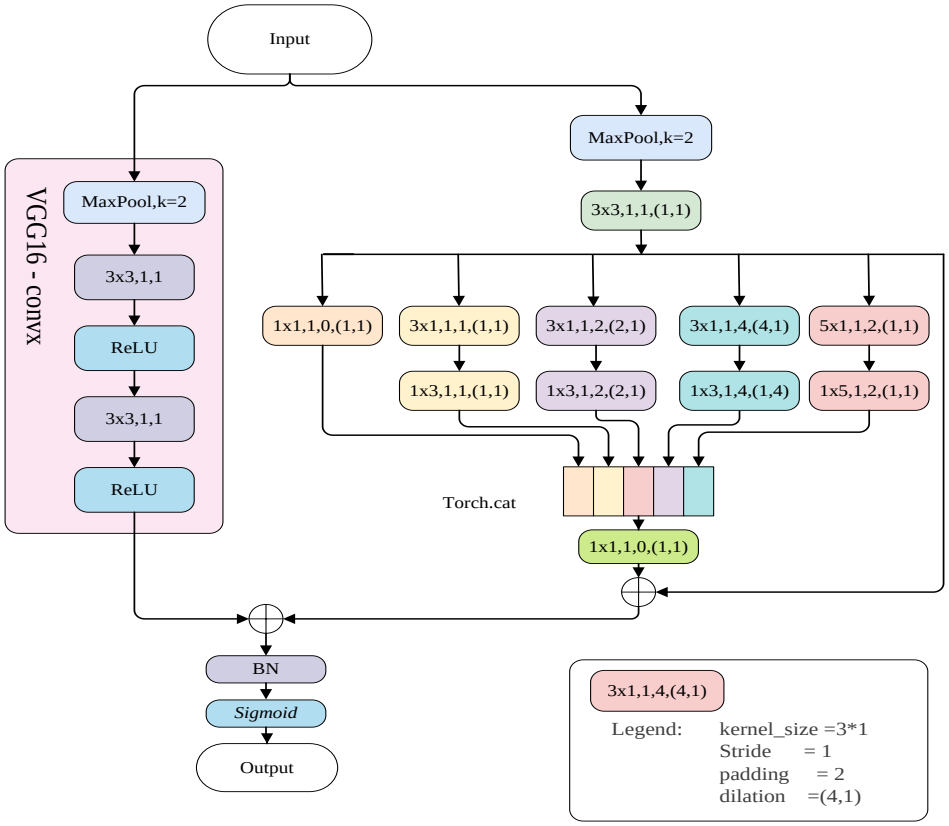


Fig. 4: The newly designed VMAC module, utilized by the proposed SSD-VUPV model
Source: created by the authors

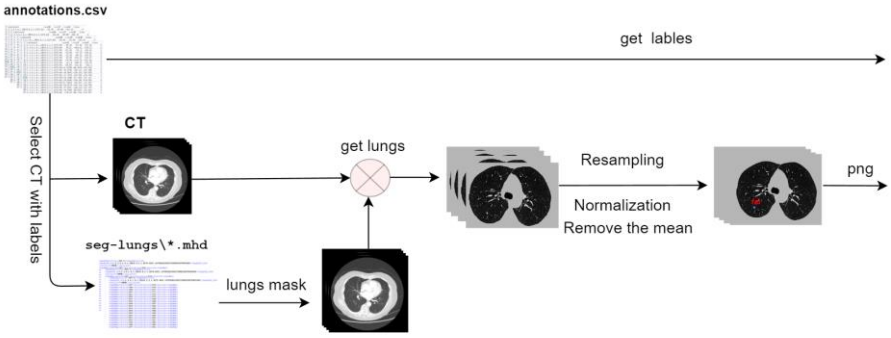


Fig. 5: The data preprocessing performed on the LUNA16 dataset
Source: created by the authors

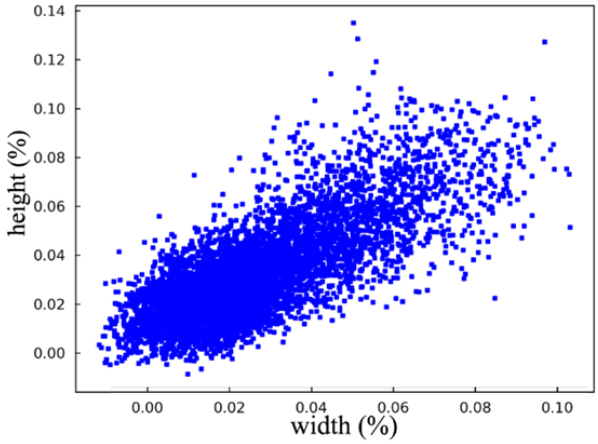


Fig. 6: The LUNA16 distribution of the proportion of the pulmonary nodule size to the entire image size
Source: created by the authors

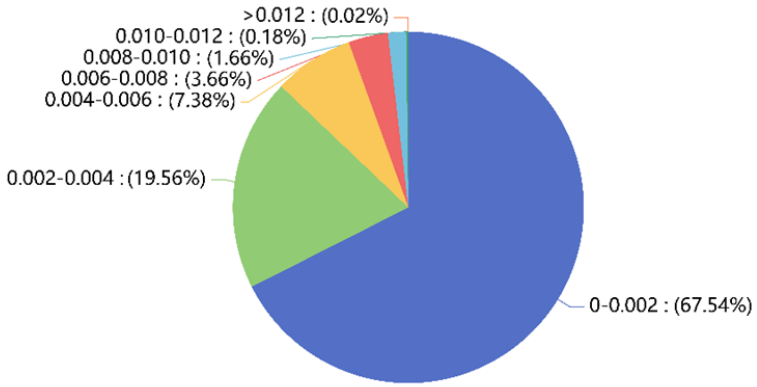
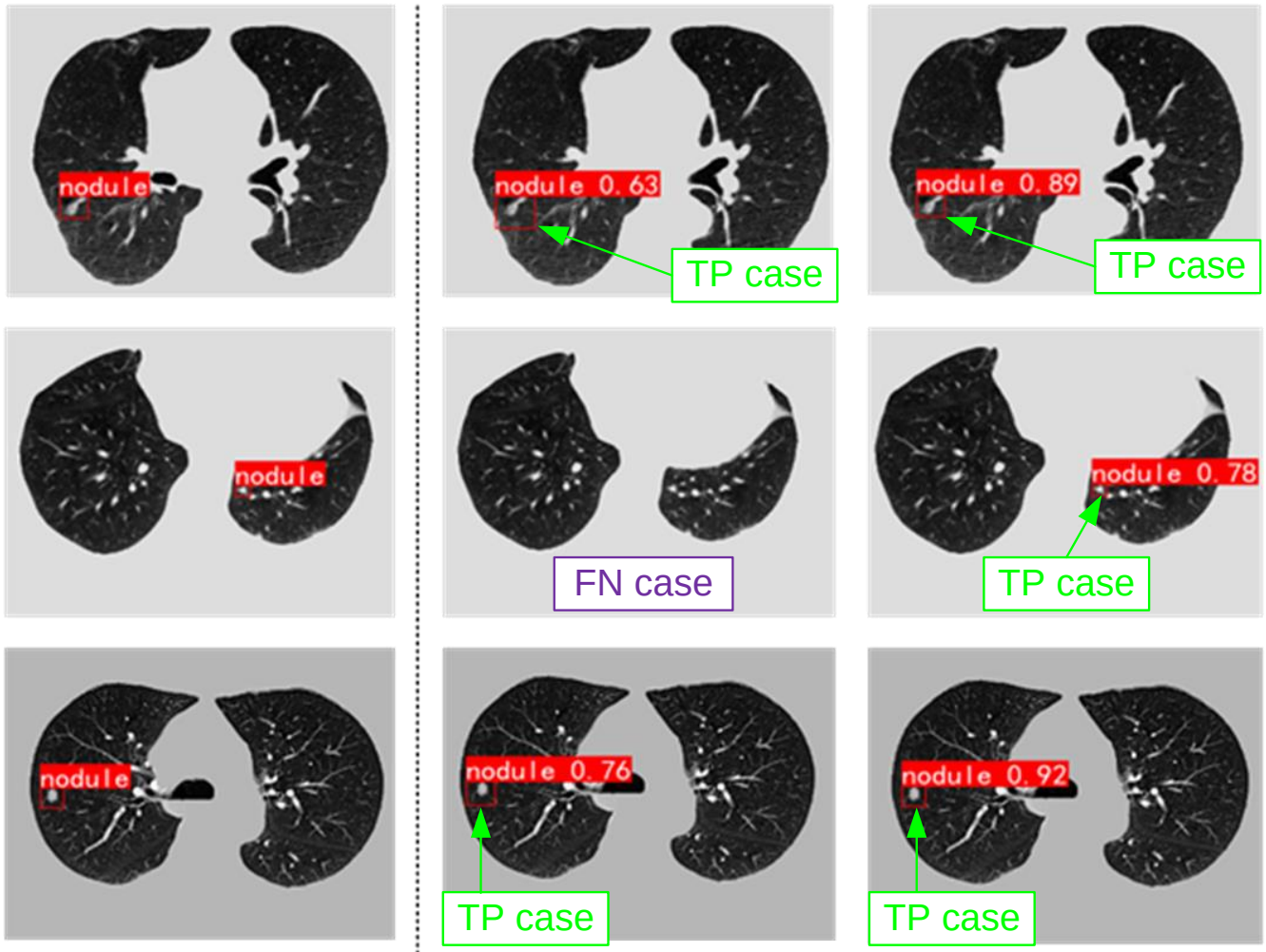


Fig. 7: The LUNA16 distribution of the ratio of the pulmonary nodule area to the entire image area
Source: created by the authors



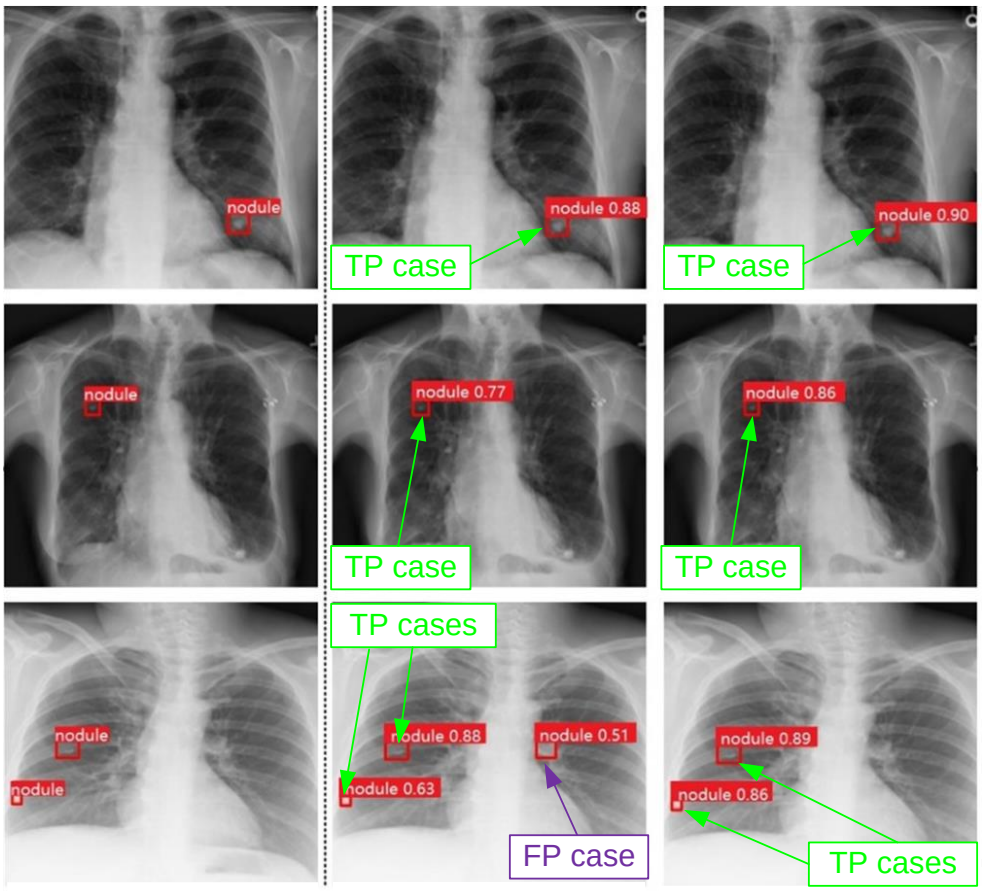
(a) Annotated original images

(b) Lung nodule detection by baseline (SSD)

(c) Lung nodule detection by proposed model (SSD-VUPV)

Fig. 8: Visualization of the detection ability of the proposed model (SSD-VUPV) vs. the baseline (SSD), based on the LUNA16 dataset

(a) Source: [28], [29]; (b)&(c) Source: created by the authors, based on [28], [29]



(a) Annotated original images

(b) Lung nodule detection by baseline (SSD)

(c) Lung nodule detection by proposed model (SSD-VUPV)

Fig. 9: Visualization of detection ability of the proposed model (SSD-VUPV) vs. baseline (SSD), based on the ChestX-ray14 dataset

(a) Source: [31]; (b)&(c) Source: created by the authors, based on [31]