Three-Dimensional Multifaceted Attention Encoder–Decoder Networks for Pulmonary Nodule Detection

Keyan Cao 1,2,*, Hangbo Tao 1 and Zhongyang Wang 1

1 School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang 110168, China
2 Liaoning Province Big Data Management and Analysis Laboratory of Urban Construction, Shenyang 110168, China
* Correspondence: caokeyan@sjzu.edu.cn

Abstract: Lung cancer is one of the most dangerous cancers in the world, and its early clinical manifestation is malignant nodules in the lungs, so nodule detection in the lungs can provide the basis for the prevention and treatment of lung cancer. In recent years, the development of neural networks has provided a new paradigm for creating computer-aided systems for pulmonary nodule detection. Currently, the mainstream pulmonary nodule detection models are based on convolutional neural networks (CNN); however, as the output of a CNN is based on a fixed-size convolutional kernel, it can lead to a model that cannot establish an effective long-range dependence and can only model local features of CT images. The self-attention block in the traditional transformer structures, although able to establish long-range dependence, are as ineffective as CNN structures in dealing with irregular lesions of nodules. To overcome these problems, this paper combines the self-attention block with the learnable regional attention block to form the multifaceted attention block, which enables the model to establish a more effective long-term dependence based on the characteristics of pulmonary nodules. And the multifaceted attention block is intermingled with the encoder–decoder structure in the CNN to propose the 3D multifaceted attention encoder–decoder network (MAED), which is able to model CT images locally while establishing effective long-term dependencies. In addition, we design a multiscale module to extract the features of pulmonary nodules at different scales and use a focal loss function to reduce the false alarm rate. We evaluated the proposed model on the large-scale public dataset LUNA16, with an average sensitivity of 89.1% across the seven predefined FPs/scan criteria. The experimental results show that the MAED model is able to simultaneously achieve efficient detection of pulmonary nodules and filtering of false positive nodules.

Keywords: pulmonary nodule detection; irregular receptive fields; encoder–decoder; computer-aided diagnosis

1. Introduction

According to data published by the World Health Organization’s Agency for Research on Cancer, in 2020 alone, 2.21 million people suffered from lung cancer and 1.8 million people died from lung cancer. Lung cancer is the second most common cancer globally and the number one cause of death. Lung cancer has no apparent complications in its early stages, so many patients are diagnosed with advanced lung cancer and miss out on the best time to cure it. The early clinical manifestation of lung cancer is a malignant nodule in the lung, so early detection and effective treatment of pulmonary nodules can significantly improve the survival rate of patients [1].

Quantitative CT imaging of the lungs can detect potential lung cancer patients better than standard chest X-ray screening [2], making CT images one of the main tools to assist physicians in diagnosing pulmonary nodules. However, radiologists must look at hundreds or even thousands of CT slices for each patient to reach an accurate conclusion. Due to the physician’s ability, work experience, and work status limitations, the results are often
not objective enough and are prone to misses and misdiagnoses. Many researchers have therefore worked to develop efficient computer-aided detection systems to achieve accurate and effective detection of pulmonary nodules in low-dose CT scans. Early computer-aided diagnostic systems, such as morphological thresholds, relied on a priori knowledge. They required manual extraction of low-level features of pulmonary nodules, making it challenging to discriminate complex nodules after specific lesions have occurred [3]. According to Wani et al. [4], AI can improve the knowledge of healthcare professionals, enabling them to spend more time on direct patient care and reduce fatigue. In recent years, deep learning of AI has demonstrated powerful learning capabilities in processing raw image tasks. The advantage of deep learning-based models is that they can learn the intrinsic representation of the data despite the absence of prior knowledge.

There are many deep learning-based approaches to achieve pulmonary nodule detection [5], primarily based on a two-stage strategy: 1. candidate pulmonary nodule detection and 2. false positive pulmonary nodule filtering. For example, Zhao et al. [6] combined multiscale blocks with Faster R-CNN blocks to construct a pulmonary nodule detection model and proposed a 3D convolution-based MSM-CNN model to filter false positive pulmonary nodules. Zhang et al. [7] incorporated a multiscale attention block into an encoder–decoder network as the backbone of a Faster R-CNN model for pulmonary nodule detection. And the 3D multiscale attention block was incorporated into the Res2Net structure for a false positive reduction in pulmonary nodules. Peng et al. [8] combined Res2Net and the channel self-attention mechanism as the backbone of the model for the detection of lungs and the false positive reduction in nodules. Although the same network structure is adopted for different tasks, the features that the two networks advance for each other do not share with each other and still need to be trained multiple times. For multi-stage network models, features are extracted from the original dataset and predicted at each stage. However, the repetitive feature extraction will bring much computational cost [9]. Therefore, many scholars have begun constructing single-stage models for detecting lung nodules. For example, Li et al. [10] designed a feature extraction network with an encoding–decoding structure. They combined it with a 3D region suggestion network (RPN) to effectively detect pulmonary nodules. The single-stage pulmonary nodule detection task, without the false positive nodule screening component, necessitates that the network has a more robust feature extraction capability to identify pulmonary nodules more accurately. Most feature extraction networks are based on convolutional neural networks (CNNs). However, the limited receptive fields of CNNs make establishing long-range dependence on the model challenging. However, long-range dependence can provide information on the location and relationship of the target across the scene, which is crucial for detecting pulmonary nodules.

Inspired by the field of natural language processing (NLP), many scholars have fused the transformer module with CNNs to overcome the inherent weakness of CNN structures that cannot establish long-range dependencies. The transformer module weights and maps each token to all other tokens through a self-attention block to obtain the dependencies between different locations [11]. Due to its unique ability to model the global context, it has been widely used in medical images [12]. However, as with traditional CNN structures, the transformer structure itself is inflexible, where each pixel point must interact with all other pixel points, inevitably introducing a portion of redundant information [13]. In particular, in the pulmonary nodule detection task, as pulmonary nodules account for a tiny proportion of the whole CT image, global modeling using the transformer introduces many lung features into the computation (nodule features account for a tiny proportion). In addition, transformer-based models can also require more data to fit than traditional CNN models, and researchers often cannot collect enough data to fit a pure transformer model [14]. In order to establish long-range dependencies with the help of transformer structures under limited datasets, often the following two approaches can be used: 1. training the model on large datasets and applying the feature extraction capabilities obtained from the training
to a new task (transfer learning) [15] and 2. using CNN structures instead of positional coding to achieve a more efficient inductive bias [16].

To address the above challenges, this paper introduces local learnable region attention (LRA) [17] into the transformer model. It combines it with traditional CNN models and 3D RPN models to propose the MAED model. LRA can form irregular receptive fields according to the specific conditions of the target and combines with traditional self-attention to form multifaceted attention, enabling the model to identify well irregularly lesioned pulmonary nodules. The multiscale fusion block (MSF) enables the model to better extract pulmonary nodule features of different sizes/diameters. The MAED model utilizes dynamic scaling with cross-entropy loss [18] to cope with the imbalance between positive and negative sample sizes. Overall, the MAED model can extract the texture information of nodules while also establishing efficient long-range dependence, enabling more accurate single-step pulmonary nodule detection.

The main points of the contribution of this paper are as follows:
1. We propose a novel pulmonary nodule detection model, which can change the size of the receptive field according to the characteristics of the detection target and can effectively deal with various irregular lesions of pulmonary nodules.
2. We use CNN structures to optimize the inherent shortcomings of the transformer-based model and enable the MAED model to effectively extract local features of nodules and establish long-range dependencies, improving the reliability and generalization of the model.
3. We evaluate the MAED model on LUNA16 data, and the results of our experiments show that our model achieved both accurate detections of pulmonary nodules and the screening of false positive nodules in 3D CT scans compared to existing methods.

2. Related Works
2.1. Three-Dimensional Pulmonary Nodule Detection

The early detection of pulmonary nodules is the best way to prevent and treat cancer. With the gradual increase in lung cancer patients in recent years, there has been much interest in using computer-aided diagnosis (CAD) systems to improve radiologists’ efficiency. Most CAD systems have two stages: 1. candidate pulmonary nodule detection and 2. false positive pulmonary nodule filtering. Most previous work has used 2D anchor-point detection networks [3]. However, 2D or multi-view 2D models can lose important information from the original image and fail to identify pulmonary nodules accurately. Therefore, the use of 2D anchor-point detection networks is often combined with 3D networks to screen for false positive nodules and achieve more accurate detection. In recent years, many scholars have constructed 3D candidate nodule detection networks that do not need to be coupled with a false positive nodule filtering network to achieve higher accuracy. For example, Khosravan et al. [19] extended the DenseNet model to a pulmonary nodule detection task. They achieved a competitive performance metric (CPM) of 0.897 on the LUNA16 dataset by combining five 3D dense blocks and the maximum pooling layer. Huang et al. [20] incorporated a one-shot aggregation block, the receptive field block, and a feature fusion scheme into the YOLOV3 model. They eliminated duplicate detection by a non-polar depressed value algorithm, achieving a CPM of 0.905 on the LUNA16 dataset. Luo et al. [9] proposed a centroid matching model (SCPM-Net) based on a 3D sphere representation. SCPM-Net, compared to traditional pulmonary nodule detection models, does not require the manual setting of anchor parameters to achieve good outcomes.

2.2. Variable Attention

The transformer model can obtain larger receptive fields than the traditional CNN model and establish a long-range dependence between pixels. Dosovitskiy et al. [21] proposed the vision transformer (ViT). By cutting the image into different patches and transforming them into different tokens through linear mapping, the dependencies between the different tokens are established with the help of the self-attention block, successfully
introducing the transformer struct into the vision domain. The transformer-based models have shown strong competitiveness in various vision tasks and are gradually being extended to medical imaging. For example, Cheng et al. [22] added a pre-trained ViT module between the encoder and decoder of the network for 2D medical image segmentation. Yang et al. [23] achieved the first multi-instance learning of pulmonary nodule features by using a transformer model to simultaneously learn multiple pulmonary nodule features from the same patient. The transformer builds long-range dependencies through self-attention while also introducing higher spatial and temporal complexity, and there are many redundant calculations in global modeling. Therefore, many scholars have attempted to construct non-global long-range dependencies. Beltagy et al. modified the traditional self-attention mechanism to propose the Longformer model [24]. In the Longformer model, each token performs local attention only for tokens in a fixed window size, and only a small number of tokens perform global attention computations. Swin-transformer divides the image into different windows, computes self-attention within each window, and devises a way to interact across windows [25]. However, most methods are based on setting a fixed-size local window, which requires the artificial setting of the window size and is poorly robust to scale-varying targets. Lin et al. [15] provided new ideas, and they extended variable attention into learnable attention by proposing the LRA block, which can dynamically generate different size perceptual fields for each target.

The remainder of this paper is organized as follows. In Section 3, we describe the complete framework of the proposed model. Section 4 describes the details of the experiment and discusses the structure of the experiment. Section 5 discusses the experimental results and future research.

3. Methods

In this section, we focus on the overall architecture of the MAED model, as shown in Figure 1, and provide details of the specific implementation of the MAED model. The following four main areas include the encoder–decoder network, learnable region attention (LRA), multiscale fusion block (MSF), and 3D RPN block for regression prediction. All abbreviations for this paper are shown in Table 1.

Table 1. List of abbreviations.

<table>
<thead>
<tr>
<th>Original Sentence</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional neural networks</td>
<td>CNN</td>
</tr>
<tr>
<td>Region suggestion network</td>
<td>RPN</td>
</tr>
<tr>
<td>Local learnable region attention</td>
<td>LRA</td>
</tr>
<tr>
<td>Multiscale fusion block</td>
<td>MSF</td>
</tr>
<tr>
<td>Computer-aided diagnosis</td>
<td>CAD</td>
</tr>
<tr>
<td>Vision transformer</td>
<td>ViT</td>
</tr>
<tr>
<td>Free-response receiver operating characteristic</td>
<td>FROC</td>
</tr>
</tbody>
</table>

Figure 1. The main structure of MAED, where the red modules are the LRA and self-attention blocks, the grey section shows the complete process of building long-range dependencies, and the blue section shows the regression prediction network.
3.1. Three-Dimensional Encoder Network

The LRA block is based on transformer implementation, and because the computational complexity of the transformer is quadratic concerning the number of tokens (sequence lengths), directly treating each pixel point of the image as a token would introduce a large amount of computation. If one follows the ViT model (segmenting the image into fixed patches and then linearly mapping them into tokens), the model loses the ability to model local information about the image in space and depth [26]. In this paper, we remedy these deficiencies by designing a set of downsampled convolutions and combining them into an encoder network that progressively encodes the input image to produce a low-resolution representation of high-level features. In this way, the feature map will be rich in local contextual information with relatively little computational effort. Each downsampling block can take the form of

\[ F(X) = Re((Re(X) \times W_{c1})) \times W_{c2} + X \]  

(1)

where \( W_{c1} \) and \( W_{c2} \) represent a 3D convolution with a convolution kernel of \( 3 \times 3 \times 3 \) (padding of 1 and step size of 1, without changing the number of channels in the feature map), which models the feature map locally at the spatial level. \( Re \) represents an activation function to the non-linear map which is ReLU. After local modeling of the feature map with two 3D convolutions, the original features are summed with the extracted features using a residual join to prevent gradient dispersion. \( W_d \) represents a 3D convolutional kernel of \( 3 \times 3 \times 3 \) with a step size of 2 and a padding of 1. This step downsamples the input feature map and expands the number of channels of the feature to twice the original. It is worth noting that we use an additional convolution with a convolution kernel of \( 3 \times 3 \times 3 \) (step size 1, padding 1) in front of the first module of the encoder to expand the input single-channel image \((x \in R^{(1\times H\times W\times D)})\) into a feature map with a channel number of 16 \((f \in R^{(16\times H\times W\times D)})\) as a way to obtain a richer feature set of the feature map.

3.2. Long-Range Dependence

As shown in Figure 2, models with different structures have different receptive fields. Because a fixed-size convolution kernel and an attention mechanism with predefined ranges cannot adequately learn spatial information between different scales [27], this paper introduces local learnable region attention (LRA), in which each feature is dynamically assigned attention. The basic idea of LRA is to introduce a learnable mask on top of the transformer global receptive field (global self-attention \( o \) mask, \( o \) is the Hadamard product) to avoid each pixel point having to interact with all other pixel points.

![Figure 2. Comparison of the perceived wildness of different structures.](image-url)
3.2.1. Global Self-Attention

The original image \((x \in R^{(1 \times H \times W \times D)})\) was passed through the 3D encoder block to obtain a highly compressed feature map \((f \in R^{(128 \times \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4})})\). In order to allow the transformer to model the spatiality and depth of the 3D image more comprehensively, we combine the spatial and depth dimensions of the image to obtain a set of feature maps \(h(h = N \times d |N = \frac{H}{4} \times \frac{W}{4} \times \frac{D}{4}, d = 128)\). After a linear mapping layer, \(h\) generates three feature vectors, \(Q \in R^{(N \times d_k)}, K \in R^{(N \times d_k)}, V \in R^{(N \times d_k)}\) \((d_k = 512)\). Then, vector \(Q\) is dotted with vector \(K\) to obtain the similarity matrix and normalized to obtain the weight matrix. Finally, the weight matrix is multiplied by the vector \(V\) to complete the global modeling of the feature map. The following equation shows global self-attention.

\[
Att_{glb}(Q, K, V) = \text{SoftMax}(\frac{QK^T}{\sqrt{d_k}})
\]

(3)

3.2.2. Local Learnable Region Attention

For each feature on the feature map, the LRA block generates a different irregular perceptual field \(\hat{R}\). Firstly, for a given vector \(Q\) and vector \(K\), two graphs \(C^1, C^2 \in R^{(N \times N)}\) of predicted coverage are generated by the four learnable parameter matrices \(W_Q^1, W_K^1, W_Q^2, W_K^2 \in R^{(d_k \times d_k)}\). The form is as follows:

\[
C^1 = \text{SoftMax}((QW^1_Q)(KW^1_K)^T)
\]

(4)

\[
C^2 = \text{SoftMax}((QW^2_Q)(KW^2_K)^T)
\]

(5)

Each value on the prediction maps \((C^1, C^2)\) represents the probability that the corresponding pixel point belongs to the mask region. To obtain a learnable three-dimensional attention map, we adjusted \(C^1\) and \(C^2\) to the four-dimensional tensor of the form \(R^{(N \times W \times H \times D)}\), i.e., for each feature the corresponding prediction coverage map will be generated. In order to obtain the exact region of the mask, for \(C^1\), a cumulative distribution function will be used to integrate each of the predicted probability maps from the lower left (bl) to the upper right (ur). The \(C^2\) integrates from the opposite direction, and the resulting integrated regions are multiplied to obtain the learnable mask regions. The formalization is as follows:

\[
F_{C_{CDF}}^{bl}(p|C^1) = \sum_{x_1 \leq x_p} \sum_{y_1 \leq y_p} \sum_{z_1 \leq z_p} C^1_i(x_i, y_i, z_i)
\]

(6)

\[
F_{C_{CDF}}^{ur}(p|C^2) = \sum_{x_1 \geq x_p} \sum_{y_1 \geq y_p} \sum_{z_1 \geq z_p} C^2_i(x_i, y_i, z_i)
\]

(7)

where \(C^1_i, C^2_i \in R^{(W \times H \times D)}, C^1_i = C^1(i, \ldots, i)\) and \(C^2_i(C^2_i) = C^2(i, \ldots, i)\) refer to any one of the feature maps, and \(p\) refers to \(C^1_i\) or \(C^2_i\) at any point on the feature map. Suppose that \(R^b_i = [F_{C_{CDF}}^{bl}(p|C^1_i)]_{p(W \times H \times D)}\) and \(R^u_i = [F_{C_{CDF}}^{ur}(p|C^2_i)]_{p(W \times H \times D)}\), i.e., assuming that \(R^b_i\) and \(R^u_i\) are the result of the calculation of the above equation. The following equation shows the formal definition of the mask region.

\[
R_i = R^b_i(oR^u_i)(C^2_i)
\]

(8)

where \(o\) denotes the Hadamard product, \(R_i \in R^{(W \times H \times D)}\).

Due to the inclusion of the softmax function in the calculation of the probability plot, the probability plot may contain a large number of zeros, and the Hadamard product can effectively extract the overlapping part of the two probability maps (mask region). However, if the predicted non-zero values are concentrated in the image’s corners, it may result in fewer overlapping parts, as shown in the first row of Figure 3. Therefore, it is necessary to integrate \(C^1\) from the top right to the bottom left again using the cumulative distribution function and \(C^2\) using the opposite direction, i.e., repeat the above steps using...
the opposite direction of integration and sum up, as shown in Figure 3. The complete masked area is as follows:

$$\tilde{R} = R^{bl}_{1}(C^1_1) \circ R^{ur}_{2}(C^2_2) + R^{ur}_{1}(C^1_1) \circ R^{bl}_{1}(C^2_2)$$  \hspace{1cm} (9)

The learnable areas of self-attention are as follows:

$$Att_{lra}(Q,K,V) = SoftMax\left(\frac{QK^T \circ \tilde{R}}{\sqrt{d_k}}V\right)$$  \hspace{1cm} (10)

Subsequent experiments demonstrated that the global perceptual information is still non-negligible, so we summed the LAR with the global self-attention to retain the global information (the global self-attention weighting is small). The complete attention struct used in this paper is as follows:

$$Att = 0.1 \ast Att_{glb} + Att_{lra}$$  \hspace{1cm} (11)

![Image of learnable region filter mechanism](image)

**Figure 3.** An example of the learnable region filter mechanism.

### 3.2.3. Feedforward Layer

We keep the feedforward layer of the transformer module and use it to transform the features. The feedforward layer consists of 2 main fully connected layers. The first fully connected layer will map the token dimension to twice the original. The second fully connected layer will then map it back to the token dimension. This process can be abstracted into a function as follows:

$$Feed\_forward(X) = Re(XW_1 + b_1)W_2 + b_2$$  \hspace{1cm} (12)

### 3.3. Multiple-Scale Fusion

The purpose of the multiscale fusion (MSF) block is to extract node features of different sizes/diameters without reducing the resolution of the feature map. This paper extracts and fuses features using different dilated rates and $1 \times 1 \times 1$ convolution. The MSF block effectively prevents the loss of features, and its main structure is as follows:

$$MSF(X) = X + down(cat(X \times W_{m1}, X \times W_{m1} \times W_{ms1}, X \times W_{m1} \times W_{ms2}))$$  \hspace{1cm} (13)

where $W_{m1}$ is a $1 \times 1 \times 1$ convolution, $W_{ms1}$ and $W_{ms2}$ are two 3D convolutions with dilation rates of 1 and 2, respectively (convolution kernel of $3 \times 3 \times 3$, step size of 1, and padding of 1). After contacting the feature channels, a bottleneck layer (a convolution kernel of $1 \times 1 \times 1$) fuses the features and reduces the model’s parameters. Finally, the original features are summed with the extracted features via a residual block to preserve the original features, as shown in Figure 4. Dilated convolutions with different dilation rates result in different receptive fields. Therefore, combining convolutions with different dilation rates can improve the feature learning ability of the network for different scales of pulmonary nodules.
3.4. Three-Dimensional Decoder Network

Because pulmonary nodules comprise a small proportion of the lung image, regressing predictions directly on a highly compressed feature map is challenging. Hence, this paper introduces a decoder block of a 3D CNN for upsampling to extend the resolution of the feature map, as shown on the right-hand side of Figure 2. Each of the upsampling modules can be abstracted as follows:

\[ F_{up}(X, X_{pre}) = \text{Cat}((X \times W_{d1}) \times W_{tra}, X_{pre}) \times W_{d2} \]  

\[ F_{ex}(X) = \text{Re}(\text{Re}(X \times W_{c1}) \times W_{c2}) + X \]

The function \( F_{up} \) amplifies the resolution of the feature map and models the combination of feature maps at the corresponding positions in the encoder using a jump connection. \( W_{d1} \) and \( W_{d2} \) represent 3D convolutions with the 1 \times 1 \times 1 kernel, which compresses the number of channels of the original feature map to half. \( W_{tra} \) is a transposed convolution with the 3 \times 3 \times 3 3D kernel, which extends the resolution of the original feature map by a factor of two. \( X_{pre} \) refers to the input of the jump connection. \( F_{ex} \) consists mainly of two 3 \times 3 \times 3 3D convolutions and residual links, which are used to further learn the feature map. The first part of the decoder has no upsampling and no jump connections, so the \( F_{ex} \) block without residual connections is used instead of the \( F_{up} \) block.

3.5. Three-Dimensional Region Generation Network

We use a 3D region generation network (RPN) to achieve the final discrimination of pulmonary nodule classes (positive and negative nodules), predict each nodule bounding box, and merge overlapping candidate boxes using non-maximal suppression. Specifically, two sets of vectors are generated for each pixel point using two convolutional layers (with a convolutional kernel of 1 \times 1 \times 1) predictions \( y_1, y_2 \), where \( y_1 \) is a one-dimensional vector indicating the probability that the pixel point is a positive nodule, and \( y_2 \) is a four-dimensional vector predicting the offset of the anchor point of the spatial location and radius prediction. Because negative nodules will be much more significant than positive ones, we use a focal loss function to match positive and negative samples. The focal loss function is defined as follows:

\[ L_{cls} = -\delta(1 - p_t)^\tau \log(p_t) \]  

In this study, when the label contains an accurate sample, \( p_t = y_1 \); otherwise, \( p_t = 1 - y_1 \). \( \delta \) is the balance factor and \( \tau \) is the focusing parameter (\( \delta = 0.5, \tau = 2 \)).

The regression loss function on the nodal bounding box is

\[ L_{reg} = \sum_k S(G_k, P_k) \]
\[ G_k = \left( \frac{x_g - x_a}{r_a}, \frac{y_g - y_a}{r_a}, \frac{z_g - z_a}{r_a}, \log \left( \frac{r_g}{r_a} \right) \right) \]  
(18)

\[ P_k = \left( \frac{x - x_a}{r_a}, \frac{y - y_a}{r_a}, \frac{z - z_a}{r_a}, \log \left( \frac{r}{r_a} \right) \right) \]  
(19)

where \( x, y, z, r \) denote the predicted offsets and \( x_a, y_a, z_a, r_a \) denotes the defined anchor point, and \( x_g, y_g, z_g, r_g \) denote the true label. \( s \) denotes the smoothed L1 regularization function.

4. Experiment

4.1. Data and Pre-Processing of Data

In this paper, the proposed model was evaluated comprehensively using the publicly available dataset LUNA16 [28], a filtered subset of the LIDC/IDRI dataset. The researchers excluded nodules with slice thicknesses greater than 2.5 mm and less than 3 mm from the LIDC/IDRI dataset and gave accurate annotation information. The LUNA16 dataset is a more suitable model for assessing the detection and segmentation of pulmonary nodules than the LIDC/IDRI dataset. In the LUNA16 dataset, there are 1186 pulmonary nodules from 888 different CT scans with a mean nodule size of 8.31 mm (standard deviation of 4.76 mm).

The original LDCT images contain many background areas besides the lung region that cannot be used directly in the experiment. This paper’s input images were finely segmented using a binary segmentation masking method (the LUNA16 dataset provided masking) to extract the complete lung tissue structure. All the pixel points were then cropped to the range \([-1200,600]\) and normalized. The LUNA16 dataset is derived from devices made by different manufacturers, and inter-device differences lead to differences in the parameters of the images. For example, the slice thickness of CT scans varies (between 0.45 and 5.0 mm); the spacing of image pixels (in millimeters) also varies. The authors of this paper, therefore, resampled the cropped area, normalized the voxel spacing to 1 mm, and finally scaled all the images to \(128 \times 128 \times 128\).

4.2. Implementation Details

The experiments in this paper are based on implementing the PyTorch framework and running on a server with four NVIDIA Tesla V100 SXM2 GPUs. All the experiments in this paper were set with a random number seed of 0, a batch size of 24, and a learning rate of 0.001 and used an SGD optimizer (momentum of 0.9 and weight decay of \(10^{-4}\)) to optimize the training process. In the training phase of the model, we used random cropping, random flipping, and scaling (in the range of 0.9 to 1.1) to enhance the data. For the training of anchors, the authors of this paper considered bounding boxes with an IoU (intersection and merging ratio) greater than 0.5 as positive examples and those with an IoU less than 0.02 as negative examples, as suggested by Li et al. [10]. The values of the anchor points were predefined as \((5,10,20)\). The model’s performance is evaluated in this paper using the free-response receiver operating characteristic (FROC) script provided with the LUNA16 dataset. The FROC measures the detector’s performance by calculating the average number of false alarms per scan (sensitivity) and has seven predefined FP/scan rates: 0.125, 0.25, 0.5, 1, 2, 4, and 8. All hyperparameters of the model are shown in Table 2.
Table 2. Hyperparameters optimization results.

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Parameter Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epoch</td>
<td>350</td>
</tr>
<tr>
<td>Batch Size</td>
<td>24</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.001</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.5</td>
</tr>
<tr>
<td>Optimizer</td>
<td>SGD</td>
</tr>
<tr>
<td>Momentum</td>
<td>0.9</td>
</tr>
<tr>
<td>Decay</td>
<td>$10^{-4}$</td>
</tr>
<tr>
<td>Anchors</td>
<td>5.0, 10.0, 20.0</td>
</tr>
<tr>
<td>Positive Examples</td>
<td>IoU &gt; 0.5</td>
</tr>
<tr>
<td>Negative Examples</td>
<td>IoI &lt; 0.02</td>
</tr>
</tbody>
</table>

4.3. Comparison With State-of-the-Art Models

In this paper, we further analyze the LUNA16 dataset to compare our method with several published state-of-the-art methods for pulmonary nodule detection. Dou et al. [29] extended the FCN and residual network, which have relatively good results in the visual tasks, into a 3D network for pulmonary nodule detection and false positive reduction, respectively. Zhu et al. [30] extended the 2D Faster R-CNN into 3D and designed a U-net-like structure as the backbone network of the model. Li et al. [10] incorporated channel attention into an encoder–decoder network to achieve pulmonary nodule detection and false positive reduction. Tang et al. [31] proposed a joint multi-tasking framework using a single network to simultaneously achieve pulmonary nodule detection, false positive reduction, and segmentation. Wang et al. [32] used a network with an FPN structure to detect pulmonary nodules and proposed an attention 3D-CNN network to reduce false positives in detection. Song et al. [33] developed an anchorless network for the detection of pulmonary nodules. All the models were evaluated on the LUNA16 dataset, but the preprocessing may differ. The comparative results of the experiments are shown in Table 3. The MAED model proposed in this paper achieved both the detection of pulmonary nodules and the reduction in false positives using only one network, and the overall performance of the model is at the level of the state-of-the-art methods.

Table 3. Comparison results with other state-of-the-art models.

<table>
<thead>
<tr>
<th>FROC</th>
<th>0.125</th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dou et al. [29]</td>
<td>0.659</td>
<td>0.745</td>
<td>0.819</td>
<td>0.865</td>
<td>0.906</td>
<td>0.933</td>
<td>0.946</td>
<td>0.839</td>
</tr>
<tr>
<td>Zhu et al. [30]</td>
<td>0.662</td>
<td>0.746</td>
<td>0.815</td>
<td>0.864</td>
<td>0.893</td>
<td>0.902</td>
<td>0.932</td>
<td>0.834</td>
</tr>
<tr>
<td>Li et al. [10]</td>
<td>0.739</td>
<td>0.803</td>
<td>0.858</td>
<td>0.888</td>
<td>0.907</td>
<td>0.916</td>
<td>0.920</td>
<td>0.862</td>
</tr>
<tr>
<td>Tang et al. [31]</td>
<td>0.708</td>
<td>0.783</td>
<td>0.857</td>
<td>0.900</td>
<td>0.943</td>
<td>0.955</td>
<td>0.963</td>
<td>0.873</td>
</tr>
<tr>
<td>Wang et al. [32]</td>
<td>0.712</td>
<td>0.802</td>
<td>0.865</td>
<td>0.901</td>
<td>0.937</td>
<td>0.946</td>
<td>0.955</td>
<td>0.874</td>
</tr>
<tr>
<td>Song et al. [33]</td>
<td>0.723</td>
<td>0.838</td>
<td>0.887</td>
<td>0.911</td>
<td>0.928</td>
<td>0.934</td>
<td>0.948</td>
<td>0.881</td>
</tr>
<tr>
<td>MAED</td>
<td>0.743</td>
<td>0.826</td>
<td>0.889</td>
<td>0.929</td>
<td>0.941</td>
<td>0.952</td>
<td>0.960</td>
<td>0.891</td>
</tr>
</tbody>
</table>

4.4. Ablation Experiments

In order to verify the effectiveness of the proposed module, we designed different network structures for the ablation experiments, and the effects of attentional mechanisms, local attention, LRA, and MSF on the models were analyzed, respectively. Specifically, we designed the following four variants of the model:

- Encoder–decoder: This model consists only of an encoder responsible for downsampling and a decoder responsible for upsampling. The intermediate attention blocks are replaced using the same number of residual blocks.
- Encoder–decoder+transformer: This model inserts four transformer modules between the encoder and decoder, i.e., the image is modeled using global self-attention.
- Encoder–Decoder + LRA: This model inserts four LRA blocks (with feedforward network layers) between the encoder and decoder.
• MAED model without MSF: This model does not use multiscale modules at the jump links of the encoder and decoder compared to the completed model proposed in this paper.

4.5. Results

Table 4 shows all the results of the ablation experiments, and the model proposed in this paper outperforms the other baseline models in nearly all metrics. Adding an attention mechanism to the encoder–decoder structure, i.e., adding long-term dependencies to the local modeling, can improve some of the model’s performance. Models that use only the regionally learnable attention mechanism use more parameters than if only global self-attention is used. However, performance improvement in the model is not significant. However, the model performance can be significantly improved when the two attention mechanisms are used together (with a slight weighting of the global self-attention). According to Figure 5, the FROC curve of the MAED model covers a larger area than all the other models in the ablation experiment, further demonstrating the superiority of the MAED model.

Table 4. The results of the ablation experiment.

<table>
<thead>
<tr>
<th>Model</th>
<th>0.125</th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder–decoder</td>
<td>0.711</td>
<td>0.791</td>
<td>0.856</td>
<td>0.887</td>
<td>0.903</td>
<td>0.918</td>
<td>0.928</td>
<td>0.856</td>
</tr>
<tr>
<td>Encoder–decoder+transformer</td>
<td>0.719</td>
<td>0.789</td>
<td>0.841</td>
<td>0.892</td>
<td>0.925</td>
<td>0.932</td>
<td>0.932</td>
<td>0.862</td>
</tr>
<tr>
<td>Encoder–Decoder + LRA</td>
<td>0.722</td>
<td>0.791</td>
<td>0.857</td>
<td>0.907</td>
<td>0.921</td>
<td>0.929</td>
<td>0.937</td>
<td>0.866</td>
</tr>
<tr>
<td>MAED model without MSF</td>
<td>0.734</td>
<td>0.821</td>
<td>0.893</td>
<td>0.917</td>
<td>0.940</td>
<td>0.950</td>
<td>0.960</td>
<td>0.885</td>
</tr>
<tr>
<td>MAED</td>
<td>0.743</td>
<td>0.826</td>
<td>0.889</td>
<td>0.929</td>
<td>0.941</td>
<td>0.952</td>
<td>0.960</td>
<td>0.891</td>
</tr>
</tbody>
</table>

Figure 5. Comparison of FROC curves for different networks.

To better understand the process of model inference and the assignment of weights by the model, we visualized the mask generated by the local learnable region attention (LRA), as shown in Figure 6. In the visualization, the white areas are the pulmonary nodules to be detected, with brighter colors in the plot indicating that the area will be given more weight in the calculation. As shown in Figure 6, the LRA block can accurately assign more weight to the area around the pulmonary nodule, allowing the model to focus more on the area around the pulmonary nodule and to establish practical long-range dependence on the area. However, the mask regions generated are not necessarily contiguous, i.e., irrelevant
regions are accounted for when establishing long-term dependence. In future work, we will continue to investigate further to establish more accurate and valid region perception fields and extend this to other clinical applications.

![Original image](image1)

![Mask](image2)

**Figure 6.** Visualization of the results of LRA-generated masks.

### 5. Conclusions

This paper proposes a multi-task deep learning algorithm, the MAED model, for pulmonary nodule detection and false positive nodule screening. We combined the self-attention module with the LRA block to form the multifaceted attention block, which enabled the model to establish irregular long-range dependencies based on the features of the nodules, making the model better able to deal with the irregular lesions of pulmonary nodules. We integrated the multifaceted self-attention block with the encoder–decoder structure in the CNN with each other, which enabled the proposed model to model the local texture features of nodules while establishing an effective long-range dependency, making the model better able to identify the fine nodules. The performance of the model was further improved using methods, such as multiscale modules and focal loss functions. We conducted extensive experiments on the LUNA16 dataset to validate that the MAED model had the effectiveness of each component, and the overall performance of the model had an average sensitivity of 89.1% at seven predefined FPs/scans. The model proposed in this paper has better performance compared to state-of-the-art single- or multi-stage detection methods.

Representativeness/racial bias will occur when the datasets used to develop AI algorithms are not sufficiently diverse to represent many different population groups and/or characteristics. Therefore, in future work, we will collect more clinical data from different regions/countries to avoid the over-concentration of data leading to representativeness/ethnicity bias of the model. In addition, we will further expand the pulmonary nodule detection model into a pulmonary nodule segmentation model to provide more help to radiologists in their diagnostic work.

**Author Contributions:** Conceptualization, K.C.; Methodology, K.C. and Z.W.; Validation, Z.W.; Data Curation, H.T. and Z.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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


5. Li, R.; Xiao, C.; Huang, Y.; Hassan, H.; Huang, B. Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: A review. *Diagnostics* 2022, 12, 298. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.