informatics

MDPI

Article

DA OMS-CNN: Dual-Attention OMS-CNN with 3D Swin Transformer for Early-Stage Lung Cancer Detection

Yadollah Zamanidoost [1,*], Matis Rivron [2], Tarek Ould-Bachir [1] and Sylvain Martel [1]

[1] Department of Computer Engineering, Polytechnique Montréal, Montreal, QC H3T 1J4, Canada; t.ould-bachir@polymtl.ca (T.O.-B.); sylvain.martel@polymtl.ca (S.M.)
[2] National Institute of Sciences, INSA Lyon, 69621 Villeurbanne, France; matis.rivron@insa-lyon.fr
* Correspondence: yadollah.zamanidoost@polymtl.ca

Abstract

Lung cancer is one of the most prevalent and deadly forms of cancer, accounting for a significant portion of cancer-related deaths worldwide. It typically originates in the lung tissues, particularly in the cells lining the airways, and early detection is crucial for improving patient survival rates. Computed tomography (CT) imaging has become a standard tool for lung cancer screening, providing detailed insights into lung structures and facilitating the early identification of cancerous nodules. In this study, an improved Faster R-CNN model is employed to detect early-stage lung cancer. To enhance the performance of Faster R-CNN, a novel dual-attention optimized multi-scale CNN (DA OMS-CNN) architecture is used to extract representative features of nodules at different sizes. Additionally, dual-attention RoIPooling (DA-RoIpooling) is applied in the classification stage to increase the model's sensitivity. In the false-positive reduction stage, a combination of multiple 3D shift window transformers (3D SwinT) is designed to reduce false-positive nodules. The proposed model was evaluated on the LUNA16 and PN9 datasets. The results demonstrate that integrating DA OMS-CNN, DA-RoIPooling, and 3D SwinT into the improved Faster R-CNN framework achieves a sensitivity of 96.93% and a CPM score of 0.911. Comprehensive experiments demonstrate that the proposed approach not only increases the sensitivity of lung cancer detection but also significantly reduces the number of false-positive nodules. Therefore, the proposed method can serve as a valuable reference for clinical applications.

Keywords: lung cancer detection; Faster R-CNN; computed tomography images; optimised multi-scale CNN; metaheuristic optimization; dual attention mechanism; 3D Swin Transformer

Academic Editor: Luc Bidaut

Received: 10 March 2025
Revised: 27 June 2025
Accepted: 3 July 2025
Published: 7 July 2025

Citation: Zamanidoost, Y.; Rivron, M.; Ould-Bachir, T.; Martel, S. DA OMS-CNN: Dual-Attention OMS-CNN with 3D Swin Transformer for Early-Stage Lung Cancer Detection. *Informatics* **2025**, *12*, 65. https://doi.org/10.3390/informatics12030065

1. Introduction

The pursuit of technological advancements in healthcare remains a continuous and pressing endeavor, especially in light of the critical need to mitigate the devastating effects of serious illnesses such as cancer [1–3]. Among the myriad forms of this disease, lung cancer represents a particularly formidable global threat, claiming countless lives with little warning. Data from the world health organization (WHO) [4] starkly illustrate the scale of this issue, with 2.21 million new lung cancer cases reported in 2020, constituting 11.4% of all cancer diagnoses worldwide. Furthermore, the estimated 1.8 million deaths attributed to lung cancer that year reaffirm its status as the primary cause of cancer-related mortality on a global level. Despite the differences in lung cancer prevalence across regions, demographics, and age groups, there is an unwavering need for early-stage detection

to improve patient outcomes. Early diagnosis is widely recognized as a critical factor in enhancing the success of treatment interventions and increasing survival rates.

In response to this urgent health challenge, medical researchers and technology experts have joined forces to explore innovative strategies that can potentially transform the approach to lung cancer diagnosis and therapy. One of the most promising areas of advancement involves the application of deep learning (DL) algorithms to the identification of lung nodules within diagnostic imaging modalities such as X-rays [3], computed tomography (CT) scans [2], and magnetic resonance imaging (MRI) [5]. In particular, the Faster R-CNN algorithm has emerged as a prominent tool for the early detection of lung cancer.

The integration of advanced technologies like Faster R-CNN into the diagnostic process marks a significant step forward in equipping healthcare professionals with powerful tools for more accurate detection and treatment of lung cancer [6]. This synergy between medical imaging and DL algorithms offers a beacon of hope in the fight against lung cancer, as the remarkable capabilities of Faster R-CNN for accurate identification of cancerous nodules open up new possibilities for early intervention. As efforts continue globally to address this pressing health issue, there is renewed optimism for a future where early lung cancer diagnosis can become a standard, potentially saving numerous lives and providing hope to those affected by this devastating disease. Faster R-CNN operates as a two-stage, region-based detection system that excels in extracting significant information from medical images, including MRIs and CT scans. Its detection methodology initiates with the generation of a comprehensive set of candidate regions, followed by classification and refinement using convolutional neural networks (CNNs).

The Faster R-CNN method is extensively utilized for detecting objects in medical imaging, particularly for identifying lung nodules in CT scans. A significant obstacle encountered when employing this technique for lung nodule detection is achieving adequate accuracy. The precise identification of small nodules is essential for the early detection of lung cancer. Although the CNNs that are part of the Faster R-CNN framework excel at feature extraction, they often overly generalize the attributes of these small nodules during the convolutional process, which can lead to less than optimal detection results.

In ref. [7], we introduced a novel architecture, OMS-CNN, which employs VGG16 as its backbone to enhance feature extraction. This architecture improves feature map representation by integrating the final layers of VGG16 and optimizing the number of merged channels to facilitate the detection of both large and small lung nodules. To further refine this process, the advanced PSF-HS optimization algorithm [8] is applied for channel selection, while the BAS optimization algorithm [9] is utilized for initializing the weights and biases of the merged layers. Additionally, an ensemble of multiple 3D CNNs is incorporated to mitigate false-positive detections. The overall framework of this approach is depicted in Figure 1. In this study, we propose an enhanced Faster R-CNN model based on DA OMS-CNN, which improves the sensitivity of early-stage lung nodule detection.

Although Faster R-CNN has demonstrated considerable success in object detection tasks within medical imaging, its effectiveness in identifying small and morphologically diverse lung nodules remains limited [10]. This shortcoming is primarily due to the constraints of traditional CNN-based feature extractors and standard RoIPooling methods, which often struggle to capture subtle spatial details and contextual variations critical for early-stage nodule detection. To overcome these challenges, we enhance the OMS-CNN architecture with a dual-attention mechanism designed to strengthen the model's capacity to emphasize both spatially significant regions and channel-specific features within the input data. Additionally, we propose a novel dual-attention RoIPooling (DA-RoIPooling) module that integrates attention into the region of interest feature extraction process.

This approach enables the model to better isolate and utilize the most salient features within each region, thereby improving its ability to discriminate true nodules from benign structures or artifacts. Collectively, these methodological advancements aim to address key limitations of existing Faster R-CNN-based approaches by improving detection sensitivity and substantially reducing false-positive rates. The main contributions of this paper diverge from the existing literature in several key aspects:

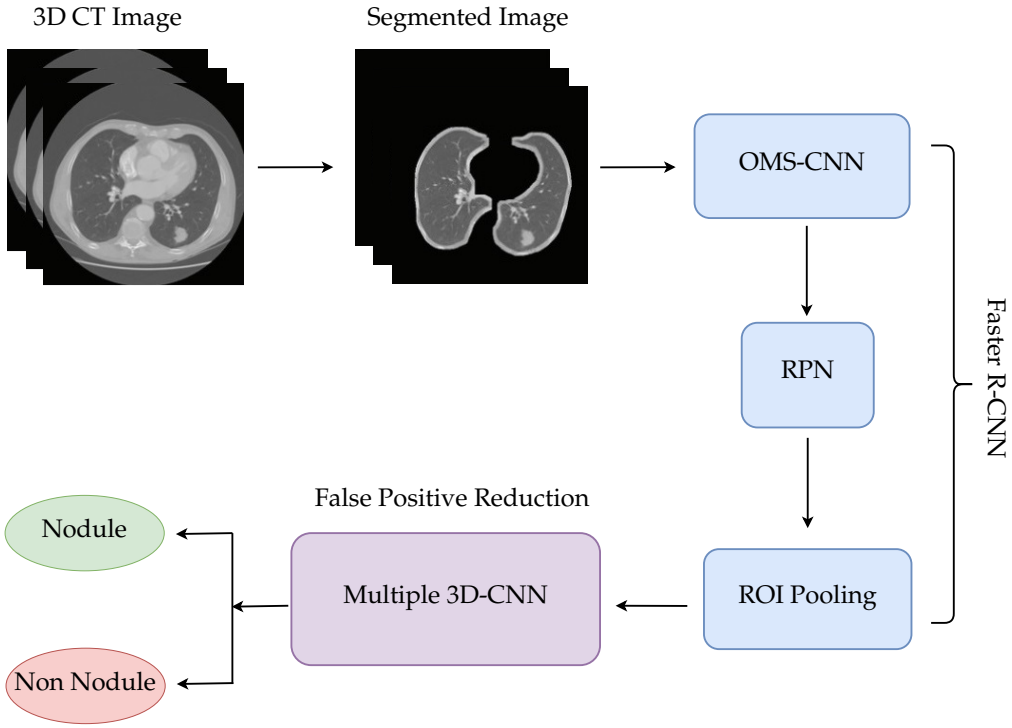

Figure 1. Overall framework of an automatic pulmonary nodule detection system based on OMS-CNN [7].

- The first contribution of this study is the integration of a dual-attention mechanism into the final layers of the OMS-CNN. The dual-attention mechanism enhances the network's ability to capture both spatial and channel-wise dependencies within the feature maps. By incorporating both spatial attention, which emphasizes important regions in the image, and channel attention, which focuses on relevant feature channels, the DA OMS-CNN achieves improved sensitivity in detecting small lung nodules. This approach ensures that critical regions and fine-grained details in the input data are highlighted, leading to more accurate and robust feature extraction.
- The second contribution is the introduction of the dual-attention RoIPooling (DA-RoIPooling) mechanism at the classification stage of the framework. DA-RoIPooling applies spatial and channel-wise attention to the pooled features, enabling the model to focus on the most relevant features within each region of interest (RoI). This dual-attention mechanism ensures that the classification network emphasizes the key characteristics of the nodules while suppressing irrelevant background information. By refining the feature representation within the RoIs, DA-RoIPooling improves the overall classification accuracy, particularly in distinguishing true nodules from false positives. This innovation significantly enhances the performance of the Faster R-CNN framework by reducing misclassifications and improving sensitivity and precision, particularly for challenging cases.
- The third contribution involves the utilization of three distinct 3D Swin Transformers for the false-positive reduction stage. This approach leverages the powerful

feature representation capabilities of the 3D Swin Transformer, which uses hierarchical feature extraction and self-attention mechanisms across spatial and temporal dimensions. By combining three separate 3D Swin Transformers, the proposed framework effectively processes volumetric data from different perspectives, ensuring a more comprehensive analysis of nodule candidates. This ensemble strategy reduces false positives by capturing subtle variations and dependencies in the 3D CT data, improving the model's ability to differentiate between true nodules and irrelevant structures. The use of 3D Swin Transformers in this stage not only enhances the overall detection accuracy but also strengthens the robustness of the proposed framework in clinical scenarios.

2. Related Works

Zamanidoost et al. [11] present a study focused on improving the detection of lung cancer nodules by enhancing feature extraction in convolutional networks. The research addresses the limitations of standard models like VGGNet and ResNet in detecting small objects, such as lung nodules, due to their feature extraction limitations. The authors propose a modified approach using the VGG16 network, known for its 3 × 3 kernels and optimal layer configuration, which can effectively capture features of small objects. Their method involves combining the feature maps from the last three layers of VGG16 to create a comprehensive representation of nodules of varying sizes. A region proposal network (RPN) is used to evaluate the proposed feature map's accuracy compared to the original VGG16. Results indicate that the proposed feature map outperforms traditional VGG16 layers in capturing nodule features and maintains higher recall stability when the number of region proposals is reduced. This approach highlights the potential benefits of optimizing feature extraction strategies in convolutional networks for lung nodule detection.

Zamanidoost et al. [7] propose an enhanced lung nodule detection method by introducing an optimized multi-scale CNN (OMS-CNN) within the Faster R-CNN framework to address the challenges of detecting nodules of varying sizes, particularly small ones. Their approach combines feature maps from the last three layers of the VGG16 architecture to create a detailed representation of nodules, which is further optimized using advanced metaheuristic algorithms, specifically the PSF-HS and BAS. These algorithms fine-tune the number of combined channels and initialize filter weights and biases, significantly enhancing feature extraction precision and efficiency. The optimized OMS-CNN effectively captures multi-scale features, improving the detection sensitivity and robustness of the model. Furthermore, a novel 3D CNN model is employed in the false-positive reduction stage, utilizing three-dimensional contextual data to refine the detection process. Experimental results on the LUNA16 and PN9 datasets demonstrate the effectiveness of the OMS-CNN in achieving higher sensitivity, reducing false positives, and achieving superior CPM scores compared to existing models, highlighting its potential for clinical application in lung nodule detection.

Tan et al. [12] proposed a multi-scale 3D CNN to improve lung nodule detection accuracy while reducing false positives. Their model integrates a 3D UNet++ architecture with a region proposal network and employs cross-layer feature fusion for enhanced feature learning. Using multiple input sizes and residual connections, the model achieves an average sensitivity of 87.3% on the LUNA16 dataset, outperforming UNet++ by 7.8% and VGG16 by 8.1%, demonstrating its effectiveness for clinical applications.

Recent studies highlight the role of attention mechanisms in improving lung nodule detection. Traditional 2D modules like SE and CBAM enhance feature extraction but are computationally expensive for 3D imaging. To address this, Almahasneh et al. [13] introduce AttentNet, a 3D fully convolutional attention mechanism that reduces computational

load while preserving feature quality. Evaluations on the LUNA16 dataset demonstrate its efficiency in candidate proposal and false-positive reduction, making it a suitable approach for 3D medical imaging.

Wu et al. [14] propose the multi-kernel driven 3D CNN (MK-3DCNN) to enhance lung nodule detection in CT scans. Their model integrates a residual encoder-decoder structure with a multi-kernel joint learning block to capture multi-scale spatial features. Additionally, a mixed pooling strategy improves feature representation. Experiments on the LUNA16 dataset show superior performance over existing methods, with further validation on the CQUCH-LND clinical dataset demonstrating its practical applicability.

Lung cancer is the leading cause of cancer-related deaths worldwide, emphasizing the need for early detection to improve survival rates. Deep learning has shown great potential in medical imaging, particularly for lung cancer identification in CT scans. Srivastava et al. [15] introduced the hybridized Faster R-CNN (HFRCNN), a two-stage model that generates and refines region proposals using a CNN. Trained on diverse datasets, HFRCNN achieves over 97% detection accuracy, outperforming many existing methods and highlighting the transformative role of deep learning in lung cancer diagnosis.

Ma et al. [16] propose TiCNet, a transformer-enhanced 3D CNN designed for early lung cancer detection. By integrating transformers with CNNs, TiCNet captures both short- and long-range dependencies, improving nodule characterization. The model incorporates attention blocks, multi-scale skip pathways, and a two-head detector to enhance sensitivity and specificity. Evaluations on LUNA16 and PN9 datasets show that TiCNet outperforms existing methods, demonstrating its potential for improving lung cancer screening.

Sun et al. [17] explored the use of the Swin Transformer model for lung cancer detection, demonstrating its potential to improve diagnostic accuracy for radiologists. Their study showed that the pre-trained Swin-B model achieved a top-1 accuracy of 82.26%, surpassing the Vision Transformer (ViT) by 2.529%. In segmentation tasks, the Swin-S model outperformed traditional methods, showing significant improvements in mean intersection over union (mIoU). This research highlights the effectiveness of pre-trained transformers in enhancing medical imaging performance, advancing reliable diagnostic tools for lung cancer detection.

These contributions illustrate the wide array of deep learning approaches that have been explored to enhance lung cancer detection, ranging from modifications of classical CNN architectures such as VGG16 and ResNet to the integration of attention mechanisms and transformer-based models. While these methods have undoubtedly advanced the field, they also exhibit several limitations. Many CNN-based models struggle to selectively focus on the most informative regions or features, thereby limiting their sensitivity, particularly for small or ambiguous nodules. Attention mechanisms and transformer models, though promising, often suffer from high computational costs or are insufficiently integrated into multi-stage detection pipelines. Moreover, few approaches effectively combine attention mechanisms at both the feature extraction and region pooling stages, or leverage 3D context in the false-positive reduction phase. These gaps highlight the need for a unified framework that integrates spatial and channel-wise attention, multi-scale feature learning, and volumetric analysis to improve both sensitivity and specificity. In response to these challenges, our work proposes a novel architecture that incorporates dual-attention-enhanced feature extraction, dual-attention RoIPooling, and a 3D Swin Transformer ensemble to provide a comprehensive solution for early-stage lung cancer detection.

3. Materials and Methods

Figure 2 illustrates an automated pulmonary nodule detection system based on the DA OMS-CNN. This system processes three-dimensional CT scan images as input and

outputs the positions of detected nodules. The implementation is designed to achieve high sensitivity in nodule detection while minimizing the average number of false positives per scan.

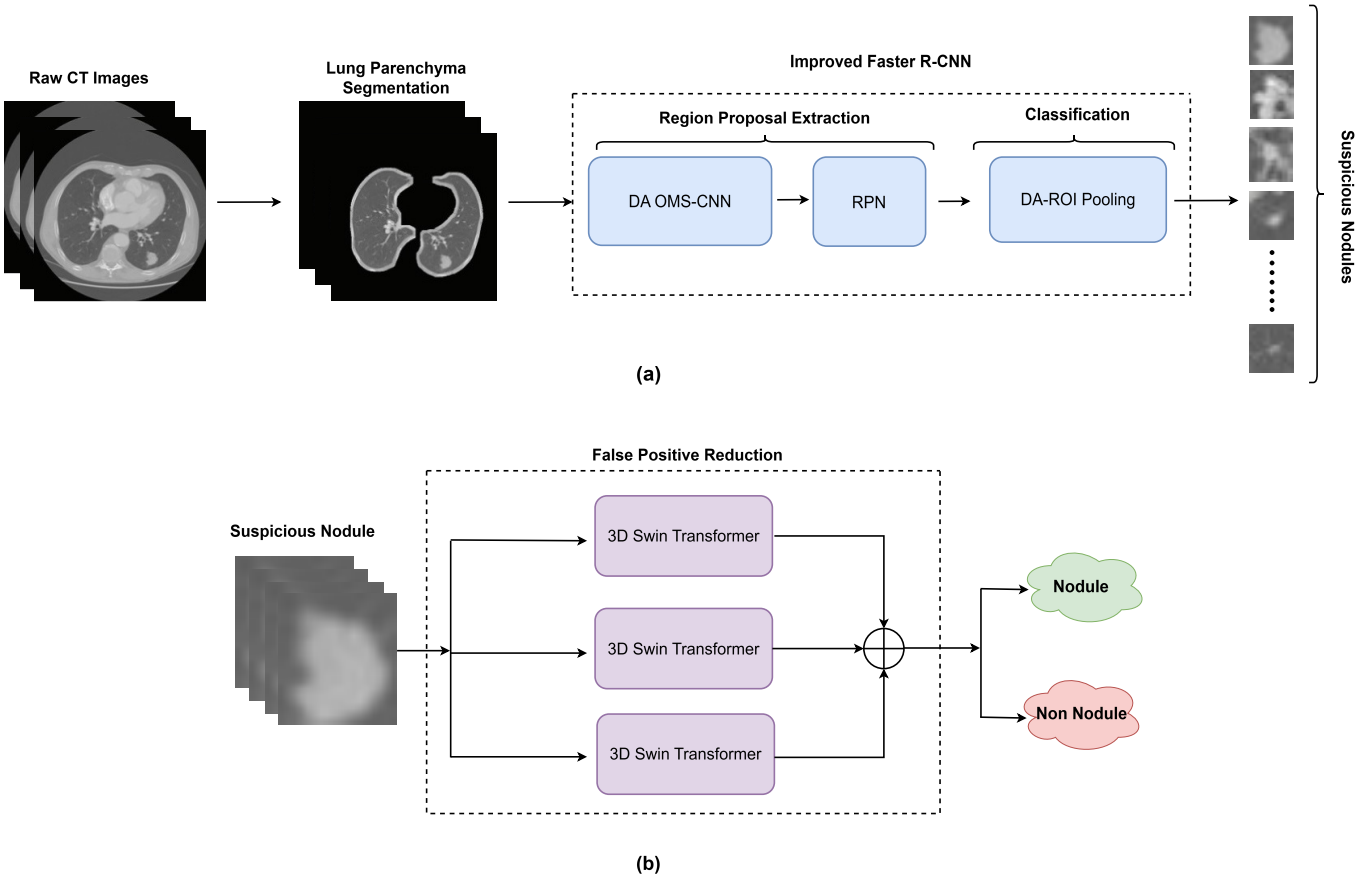

Figure 2. Overall framework of an early-stage lung cancer detection system based on DA OMS-CNN: (**a**) the proposed lung nodule detection framework; (**b**) the proposed false-positive reduction framework.

The process begins with image preprocessing, where CT scans are segmented and normalized to enhance data quality. Feature extraction is performed using the DA OMS-CNN, which incorporates spatial and channel attention mechanisms to focus on discriminative regions within the scans. Subsequently, the RPN identifies candidate regions of interest (RoIs) that may contain lung nodules. To refine these candidate regions, the DA-RoI Pooling mechanism is employed, ensuring that the most relevant features are extracted for further analysis. The refined RoIs are then processed through the stack of 3D SwinT blocks, which are used for false-positive reduction. These blocks effectively capture both global and local dependencies in the 3D space of CT scans. The outputs of the 3D SwinT networks are aggregated to deliver accurate predictions, enabling the robust identification of pulmonary nodules.

3.1. Dataset and Preprocessing

To develop and evaluate the proposed framework for lung nodule detection, two datasets were utilized: the LUNA16 [18] and PN9 [19] datasets. These datasets were chosen due to their high-quality annotations and complementary characteristics. The LUNA16 dataset was used for training, validation, and testing purposes, while the PN9 dataset was employed to evaluate the model's generalization capability. This two-pronged approach ensures that the model not only performs well on the training data but also generalizes effectively to unseen data from different clinical sources.

3.1.1. LUNA16

The LUNA16 dataset, derived from the LIDC-IDRI, is a widely used benchmark dataset in lung disease research. It consists of 888 low-dose thoracic CT scans with detailed annotations by multiple expert radiologists. The dataset includes nodules with varying sizes, shapes, and malignancy probabilities, offering a diverse set of examples for model training and evaluation. Each nodule is annotated with descriptors such as size (ranging from 3 mm to 30 mm) and shape (e.g., round, irregular, lobulated), providing a comprehensive representation of nodular characteristics. The dataset's thin-slice CT scans, with slice thickness ranging from 0.4 mm to 2.5 mm and pixel spacing between 0.310 mm and 1.091 mm, ensure high resolution for precise nodule detection.

For this study, the dataset was divided into three subsets: 70% for training (622 scans), 10% for validation (88 scans), and 20% for testing (178 scans), ensuring a balanced distribution for robust performance evaluation. Each CT scan, stored in DICOM format, comprises 100 to 500 axial slices with a resolution of 512×512 pixels. To prepare the data for input into the model, we extracted three contiguous slices for each axial slice and stacked them to form a three-channel image. These were then resized to a uniform shape of $800 \times 800 \times 3$ to ensure consistency across inputs and facilitate multi-scale feature learning during training.

3.1.2. PN9

The PN9 dataset serves as a benchmark for assessing the generalization capabilities of the proposed lung nodule detection framework. This dataset comprises CT scans collected from two major hospitals, representing diverse clinical scenarios such as outpatient visits, hospitalizations, and physical examinations. The scans were acquired over the period from 2015 to 2019, ensuring a wide temporal distribution. To guarantee data quality, the initial CT images underwent a meticulous validation process, focusing on compliance with DICOM standards. Scans containing significant respiratory motion artifacts or other disruptive interferences were excluded. Additionally, all sensitive patient information embedded within the DICOM headers—such as patient identifiers, institutional details, and referring physician names—was securely anonymized through data masking techniques.

Pulmonary nodules in the PN9 dataset were annotated using a two-step process. In the first stage, each CT scan was independently reviewed by a physician, who generated an initial medical report detailing the type, size, and approximate location of detected nodules. These reports were then cross-validated by a second physician to ensure accuracy. In the second stage, nodules were annotated in a detailed, slice-by-slice manner, with physicians referencing the corresponding medical reports to ensure consistency. For each nodule, bounding boxes and classification labels were stored in structured XML files. The nodules were categorized into nine distinct groups based on their size and type, following established medical standards. To enhance the reliability of these annotations, a second physician reviewed the outputs, and any discrepancies were resolved collaboratively.

3.1.3. Data Augmentation

In scenarios where data imbalance poses a challenge, augmentation becomes an essential strategy to enhance dataset diversity and improve model performance. Given the inherent imbalance in our dataset, we employed manual augmentation techniques to address this issue effectively. By rotating images in multiple directions and generating additional variations from different angles [20], we created a more diverse representation of the original data. These transformations mitigate the class imbalance problem and ensure a more robust learning process. Additionally, advanced augmentation methods, such as zooming in and out, applying various shear ranges, and flipping images, were utilized

to further enrich the dataset. These techniques not only introduce variability but also enable the model to interpret data from multiple perspectives, ultimately improving its generalization capabilities.

3.2. Lung Parenchyma Segmentation

Accurate segmentation of the lung parenchyma from chest CT scans is a critical preprocessing step for isolating relevant anatomical structures and enhancing the precision of nodule detection. In our approach, this task is accomplished through a series of image processing techniques, starting with intensity normalization and binarization.

Raw CT images are first clipped to a predefined Hounsfield unit (HU) range of $[-1000, 400]$, which effectively captures the range of lung tissue densities. The pixel intensities are then normalized to the interval $[0, 1]$ using the following linear transformation:

$$I_{\text{norm}}(x, y) = \frac{\min(\max(I(x, y), -1000), 400) + 1000}{1400} \tag{1}$$

where $I(x, y)$ represents the original intensity at pixel (x, y), and I_{norm} is the normalized value. This scaling ensures that lung tissue contrasts are preserved while suppressing irrelevant high-density regions such as bone.

To isolate the internal thoracic region, a global threshold T is computed as the mean of all normalized pixel intensities:

$$T = \frac{1}{N} \sum_{x=1}^{H} \sum_{y=1}^{W} I_{\text{norm}}(x, y) \tag{2}$$

Pixels with values below T are set to 1 (foreground), and others to 0 (background), generating an initial binary lung mask $B(x, y)$:

$$B(x, y) = \begin{cases} 1, & \text{if } I_{\text{norm}}(x, y) < T \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

To remove irrelevant components such as bones and air outside the lungs, a four-connected component labeling algorithm is applied. Only the two largest connected regions (corresponding to left and right lungs) are retained. All other regions are discarded as noise.

Morphological operations are used to enhance the binary mask. A morphological opening operation removes small artifacts using a circular structuring element, and a morphological closing operation is then used to fill small gaps near lung boundaries. Internal voids and gaps within the segmented lung areas are removed using a hole-filling operation [21].

The final binary mask provides a clean, contiguous representation of the lung parenchyma. This mask is subsequently used to crop the original CT images and suppress non-lung regions, ensuring that subsequent processing stages, such as nodule detection and classification, operate exclusively within the anatomically relevant domain.

3.3. Lung Nodule Detection

Detecting lung nodules, particularly small ones, is a complex challenge due to their subtle appearance and varying sizes. To address this, we developed an improved Faster R-CNN framework, enhanced with several innovative techniques designed to optimize sensitivity and accuracy for small nodule detection. Our method integrates a DA OMS-CNN to better capture multiscale features, an advanced RPN for generating accurate region proposals, and a DA-RoIPooling mechanism to enhance the classification process.

3.3.1. Dual-Attention Optimized Multi-Scale CNN (DA OMS-CNN)

Choosing the right feature extraction architecture plays a crucial role in determining the effectiveness of modern convolutional neural networks for detecting lung nodules. Several architectures, such as DenseNet, VGGNet, and ResNet, are widely used due to their ability to extract object features from images with remarkable precision. Despite their strengths, these models often struggle with accurately identifying small lung nodules. Among these architectures, VGG16 stands out for its compact 3×3 convolutional kernels and well-optimized layers, which enhance its capability to detect small nodules with improved precision. In ref. [11], the final three layers of the VGG16 network are merged to create feature maps, enabling the extraction of features from nodules of varying sizes, with a particular emphasis on smaller nodules.

This study proposes a dual-attention OMS-CNN designed for optimal RoI extraction, as illustrated in Figure 3. At this stage, fully convolutional dual-attention blocks are integrated to enhance the network's ability to focus on critical features across both channel and spatial dimensions. To achieve this, the dual-attention mechanism is incorporated at three key stages of the architecture: (1) the fourth layer of the VGG16 backbone, (2) the fifth layer of the VGG16 backbone, and (3) the concatenated feature map layer, where outputs from the last three convolutional layers are fused. The spatial attention component dynamically highlights spatial regions of interest by weighting feature maps based on their positional importance. Simultaneously, the channel attention component amplifies feature maps that contain the most discriminative information for nodule detection. This mechanism strengthens the model's focus on small nodules that may otherwise be overlooked, particularly in high-dimensional feature spaces.

To tailor the convolutional capacity of the network to different nodule scales, we define two sets of kernel configurations—$[N_S, K_S, M_S]$ for small nodules and $[N_L, K_L, M_L]$ for large nodules—corresponding to the number of output channels in the final three convolutional layers. The optimal values for these configurations are not selected heuristically but are instead determined through a principled optimization process. Specifically, we utilize the advanced parameter-setting-free harmony search (PSF-HS) algorithm [8], which formulates the search for kernel parameters as a global optimization problem. In this context, each candidate configuration is treated as a "harmony" whose fitness is evaluated based on the model's sensitivity in detecting annotated nodules within the training set. Through adaptive control of exploration and exploitation, PSF-HS iteratively refines candidate solutions and converges to high-performing configurations. This optimization strategy eliminates the need for manual tuning and improves the generalizability of the learned representations to nodules of varying shapes and sizes. For additional technical details and mathematical formulations, readers are referred to our earlier work [7]. Furthermore, to improve convergence stability and initialization quality, we adopt the BAS optimization algorithm [9], which replaces conventional random initialization for the convolutional layer filters by providing a more robust search mechanism for optimal weights and biases.

The dual-attention blocks are added only in the deeper layers of the VGG16 backbone and the concatenated feature map layer for specific reasons. In the lower layers of the network, the extracted features are low-level representations, primarily capturing general patterns such as edges, textures, and basic shapes. Applying attention mechanisms at these stages could lead to a loss of critical general-purpose information needed for constructing high-level features. Instead, the dual attention is integrated into the deeper layers (fourth and fifth) of the backbone where the feature maps are more abstract, containing high-level semantic information critical for identifying lung nodules. These layers are better suited for attention mechanisms as they focus on more meaningful regions in the image.

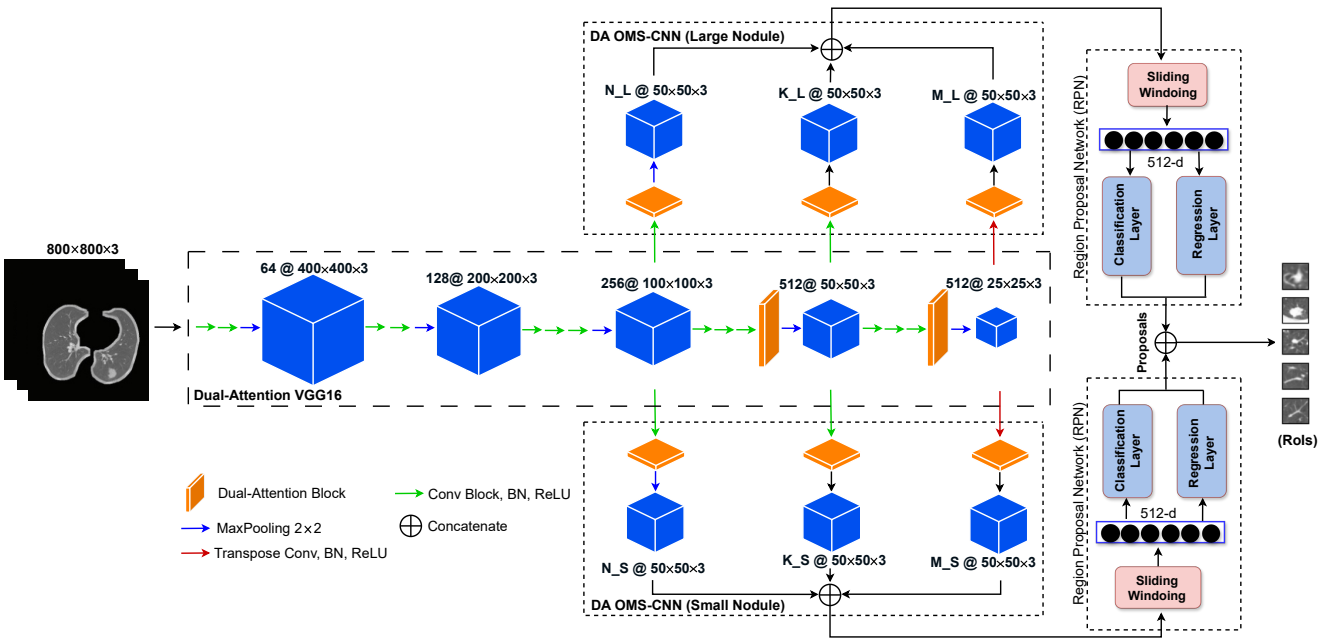

Figure 3. Overall framework of dual-attention OMS-CNN model (DA OMS-CNN).

Additionally, the concatenated feature map layer combines outputs from the last three convolutional layers, offering a multi-scale feature representation that captures nodules of varying sizes. By applying dual attention at this stage, the network selectively emphasizes the most relevant features across scales, improving its sensitivity to small nodules. Incorporating dual attention into this layer enhances the hierarchical understanding of multi-scale features while suppressing redundant or irrelevant information. This approach ensures that the network retains the ability to detect small and subtle nodules with high precision.

By avoiding the application of dual attention in the lower layers, the architecture balances the need for preserving low-level feature diversity with the refinement of high-level features. This strategic design significantly enhances the network's ability to detect nodules of varying sizes while maintaining robustness against background noise and irrelevant details, thus improving the overall sensitivity and performance of the system.

The dual-attention mechanism is designed with a sequential structure where channel attention is applied first, followed by spatial attention, as illustrated in Figure 4. This order ensures an optimal refinement of feature representations by addressing the "what" and "where" aspects of attention hierarchically [22]. By applying channel attention first, the model identifies and amplifies the most informative feature channels, effectively enhancing the global semantic understanding of the input. This step prioritizes features that are most relevant for identifying lung nodules, reducing noise at the channel level. The channel attention mechanism begins by extracting global information from the input feature map using both average pooling and max pooling operations. These operations yield two distinct descriptors, denoted as χ^{C}_{AUG} and χ^{C}_{Max}. These descriptors are then processed through a scale network, which generates a channel attention map, represented as $M_C \in \mathbb{R}^{C/2G\times 1\times 1}$. The channel attention map is subsequently used to modulate the input feature map χ, enabling element-wise summation with the corresponding sub-feature. Following this, average pooling and max pooling operations are applied to both branches of each sub-feature χ_K. The resulting feature vectors are combined using element-wise summation, producing the final output $\chi_C \in \mathbb{R}^{C/2G\times 1\times 1}$. The process can be mathematically expressed as:

$$M_C(\chi) = \mathrm{MLP}(\mathrm{AVG}(\chi)) + \mathrm{MLP}(\mathrm{MAXPool}(\chi)) \tag{4}$$

$$\chi = \text{MLP}(\chi_{K1} + \chi_{K2}) + \text{MAXPool}(\chi_{K1} + \chi_{K2}) + M_C(\chi) \tag{5}$$

To complement the channel attention mechanism, a compact feature representation is created to enable precise and adaptive selection. This is achieved using a straightforward gating mechanism with a sigmoid activation function. The final output of the channel attention is computed as:

$$\chi_C = \delta(F_C(\chi)) \cdot \chi_{K1} = \delta(W_C\chi + b_C) \cdot \chi_{K1} \tag{6}$$

Subsequently, the output of the channel attention block is passed to the spatial attention block, which focuses on localizing the critical regions of interest within the feature maps. This sequential process ensures that spatial attention operates on already refined feature maps, making it more effective at highlighting the precise locations of nodules. To compute spatial attention, we apply group normalization (GN) to the χ_{K1} and χ_{K2} branches. This approach reduces computational complexity while ensuring that spatial information is effectively utilized, providing more accurate data to the feature extraction network. The calculation for spatial attention is expressed as:

$$\chi_S = \delta(W_S \cdot (\text{GN}(\chi_{K2}) + \text{GN}(\chi_{K1})) + b_S) \cdot \chi_{K2} \tag{7}$$

Here, W_S and b_S are parameters with a shape of $\mathbb{R}^{C/2G \times 1 \times 1}$. The χ_{K1} and χ_{K2} branches are subsequently combined to align the number of channels with the input dimensions. This integration allows spatial attention to improve the representation of the feature map effectively.

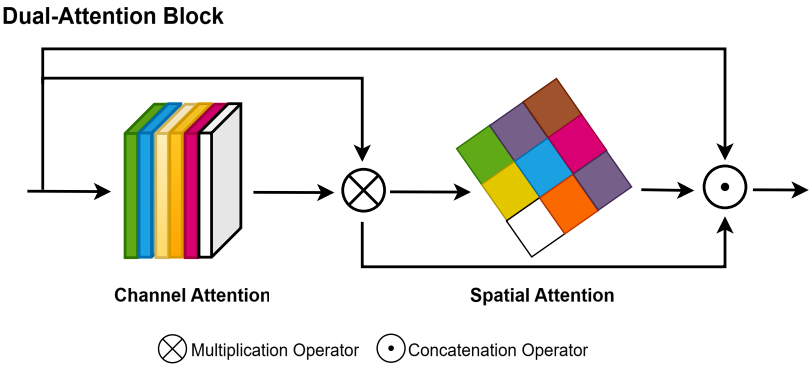

Figure 4. Proposed dual-attention block.

By separating these two stages and processing channel importance before spatial localization, the network achieves a better balance between global feature importance and local feature refinement, improving detection accuracy and robustness.

3.3.2. Region Proposal Network (RPN)

The RPN processes an input image and generates a set of rectangular proposals, each associated with an objectness score. The RPN is implemented as a fully convolutional network, designed to operate efficiently on feature maps produced by the last convolutional layer of the feature extraction network. A small network, which slides over the input feature map using a 3 × 3 spatial window, serves as the core of the RPN. Each sliding window extracts a feature vector (512 dimensions in the case of DA OMS-CNN), which is then passed through a box-classification layer to predict objectness scores and a box-regression layer to estimate the bounding box coordinates. To address the detection of both small and large lung nodules, two distinct RPNs are utilized within the framework. These RPNs are specifically designed to leverage different perspectives and extract complementary

information, which enhances the overall proposal generation process. The two networks are integrated with the DA OMS-CNN backbone, with one RPN tailored for small nodules and the other for large nodules (Figure 3). These networks operate on feature maps of the same dimensions, ensuring seamless integration with the backbone architecture. To accommodate the varying sizes of lung nodules, seven anchor boxes with different scales are employed: 4×4, 6×6, 10×10, 16×16, 22×22, and 32×32, as in [7]. This multi-scale anchor design is particularly effective in capturing nodules of diverse sizes, enabling the framework to improve detection accuracy for both small and large nodules. With these definitions, the multi-task loss function for an image is expressed as:

$$L(p_i, t_i, p_{kj}, t_{kj}) = \sum_i L_1(p_i, t_i) + \sum_{k=1}^{2} \sum_j L_2(p_{kj}, t_{kj}) \tag{8}$$

Here, L_1 and L_2 are defined as:

$$L_1(p_i, t_i) = L_{\text{cls}}(p_i, p_i^*) + \lambda \times p_i^* \times L_{\text{reg}}(t_i, t_i^*) \tag{9}$$

$$L_2(p_{kj}, t_{kj}) = \frac{1}{N_{\text{cls}}} \times L_{\text{cls}}(p_j, p_j^*) + \frac{\lambda}{N_{\text{reg}}} \times p_j^* \times L_{\text{reg}}(t_{kj}, t_j^*) \tag{10}$$

The regression loss is defined as:

$$L_{\text{reg}}(t_i, t_i^*) = R(t_i - t_i^*) \tag{11}$$

$$R(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \tag{12}$$

In these equations:

- i represents the index of the proposals generated by the region proposal networks.
- j identifies an anchor selected.
- k refers to one of the two region proposal networks.
- p_i is the predicted probability of proposal i being a nodule.
- p_i^* is the ground truth label, where $p_i^* = 1$ if the proposal is positive, otherwise $p_i^* = 0$.
- t_i and t_i^* are the predicted and ground truth bounding box regression parameters, respectively.
- L_{cls} is a binary cross-entropy loss.
- L_{reg} represents the regression loss.
- λ is a balancing factor between the classification and regression losses.
- N_{cls} and N_{reg} are normalization terms for classification and regression, respectively.
- R is the smooth L_1 function.

3.3.3. Classification Stage

After obtaining the RoIs predicted by the RPN and removing duplicates, a deep convolutional neural network (DCNN) is employed to classify each RoI, determining whether it corresponds to a nodule or not. The RPN regression layer generates candidate nodule positions, specifying the center coordinates as well as the width and height (W, H) of each RoI. These values are used to extract patches from the feature map, which serve as input to the classification network. The RPN classification layer provides a probability score for each patch, ranging between 0 and 1. Patches with scores exceeding a threshold of

0.5 are considered nodule candidates and forwarded to the classification stage for further analysis [23].

In the proposed method, we introduce a dual-attention mechanism after the RPN stage and before the fully connected layers in the RoIPooling structure, as shown in Figure 5. An RoIPooling layer is employed to project each RoI onto a smaller feature map with a predetermined spatial dimension of $W \times H$ (specifically, 7×7 as outlined in this paper). The RoIPooling process involves dividing the RoI into a grid of sub-windows measuring $W \times H$ and performing max-pooling within each sub-window, resulting in values being mapped to their corresponding output grid cells. This pooling operation is carried out independently across each feature map channel, akin to standard max pooling procedures.

Following the RoI pooling operation, the dual-attention mechanism is integrated to enhance the extracted feature representations. The channel attention block selectively emphasizes informative channels while suppressing less relevant ones, ensuring that critical features for nodule classification are highlighted. The output of the channel attention block is then passed through the spatial attention block, which focuses on relevant spatial regions within each feature map. This combination allows the network to refine RoI feature maps by simultaneously considering channel-level and spatial-level dependencies. By applying dual attention at this stage, we aim to better capture subtle and discriminative features critical for accurate classification, especially in challenging cases.

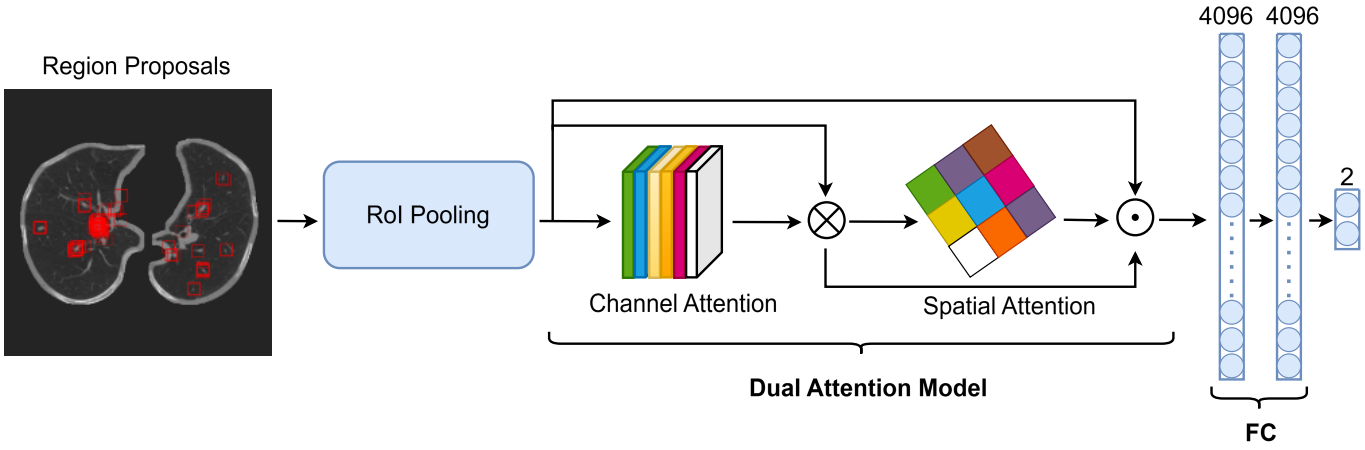

Figure 5. Overall framework of classification stage (DA RoIPooling).

After dual-attention processing, a fully connected network comprising two 4096-dimensional fully connected layers is employed to transform the fixed-size feature map into a feature vector. Finally, a binary classifier predicts confidence scores for potential candidates. The training of the classification model utilizes CrossEntropyLoss as the loss function to optimize the network. This enhanced architecture aims to reduce false positives and improve the overall sensitivity and specificity of the nodule detection pipeline.

3.4. False Positive Reduction

In the false-positive reduction phase, we employ a sequence of 3D Swin Transformer models to enhance classification accuracy and reduce false positives. The pipeline processes 3D image patches, where each patch is passed through multiple trained 3D SwinT models, as shown in Figure 6. The outputs from these models are combined using a voting mechanism to determine the final classification as either "Nodule" or "Non-Nodule." This approach leverages the hierarchical structure and self-attention mechanism of the 3D SwinT, enabling the extraction of both local and global features from volumetric data for robust decision making.

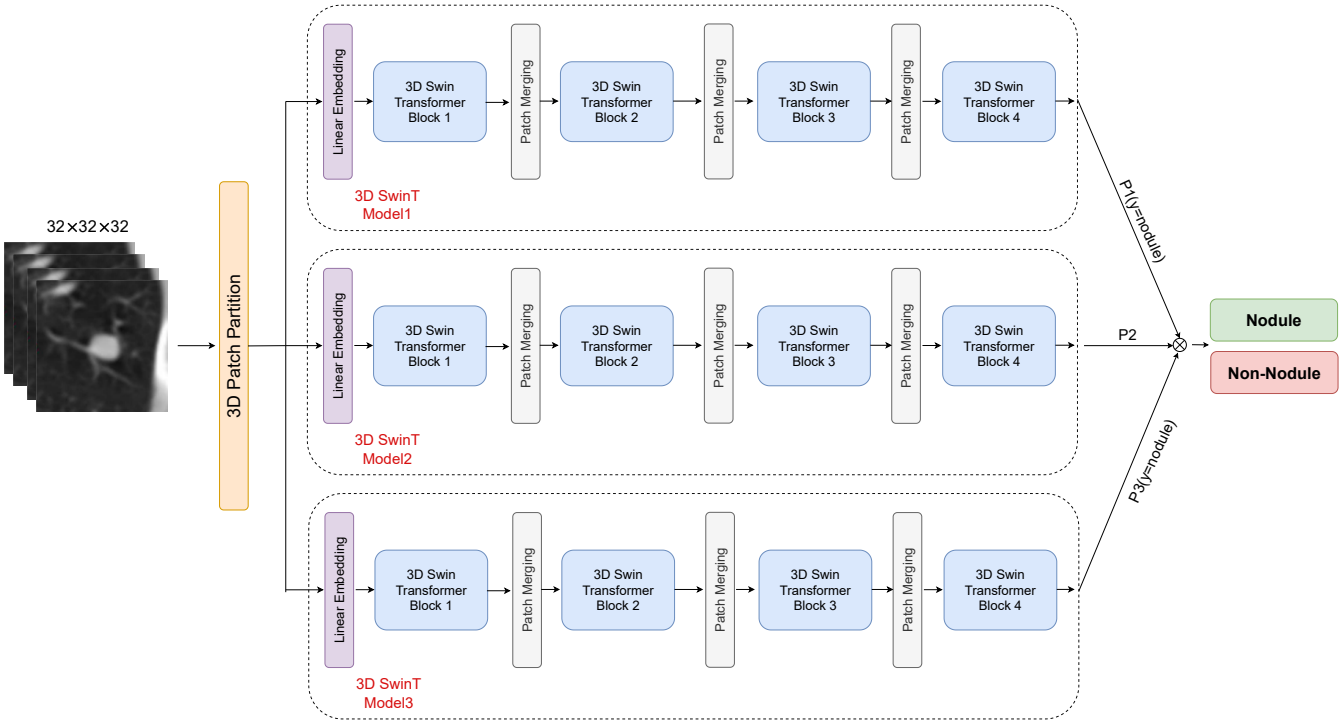

Figure 6. The structure of the proposed false-positive reduction model.

One of the main challenges in object detection is the overwhelming number of negative samples, which dominate the total loss. Many of these samples are relatively easy to classify, highlighting the importance of hard sample mining to improve performance. Following this concept, the training data is curated to emphasize more difficult samples, which persist through subsequent training iterations, enhancing the classification accuracy of each individual model.

The 3D SwinT, chosen for this phase, is designed to process volumetric data effectively by leveraging shifted window-based multi-head self-attention mechanisms. The hierarchical structure of the transformer enables the model to capture both global and local spatial relationships within the patches, offering improved performance compared to traditional convolutional models. Each model is initialized and trained independently using a specific subset of the data, focusing on misclassified samples from previous iterations [7]. Initially, the first subset is used to train Model1. Misclassified samples from both Model1 and the second subset are subsequently used to train Model2. Similarly, Model3 is trained using misclassified samples from the first two models and the third subset. This iterative training process ensures that challenging examples are emphasized, enabling the models to learn robust and discriminative features. The models are fine-tuned during successive iterations, refining their weight parameters to enhance their ability to classify difficult samples.

This iterative approach, combined with the 3D SwinT's ability to effectively capture spatial dependencies and represent complex patterns, significantly improves the classification accuracy of the false-positive reduction system. To further enhance the performance of the false-positive reduction phase, we employ various patch augmentation techniques and leverage the advanced hierarchical design of the 3D SwinT. These approaches are discussed in detail in the subsequent subsections.

3D Swin Transformer

The Swin Transformer (SwinT) is a hierarchical transformer that efficiently generates multiscale feature maps by integrating neighboring patches and employing a window partition mechanism. This approach ensures linear computational complexity relative to

image size, which is particularly advantageous for dense prediction tasks and processing high-resolution images. To adapt this architecture for the 3D characteristics of CT images, we extend SwinT into a 3D structure (3D SwinT), enabling it to capture detailed spatial and volumetric information. The architecture of 3D SwinT, illustrated in Figure 7, differs from the standard SwinT in several key aspects:

- CT images are represented as $H \times W \times D$, where D refers to the depth, and H and W denote the image's height and width, respectively.
- The patch partitioning process in SwinT divides the input into $(H/4) \times (W/4)$ patches, each sized 4×4. In contrast, 3D SwinT utilizes 3D cubes of size $4 \times 4 \times 4$, producing $(H/4) \times (W/4) \times (D/4)$ patches. These patches, with a feature dimension of 64, are projected into an arbitrary dimension C via a linear embedding layer. Following this, the neighboring patches are combined during the patch merging stage, where the spatial and depth resolution decrease progressively (4, 8, 16, 32).
- The main distinction between the SwinT and 3D SwinT blocks lies in the multi-head self-attention mechanism. For 3D SwinT, the window-based multi-head self-attention (W-MSA) is extended into a 3D version (3D W-MSA), incorporating the volumetric information. This is achieved using 3D windows sized $P \times M \times M$, where P represents the depth dimension, instead of the 2D $M \times M$ windows used in SwinT. Additionally, the window shifting mechanism in 3D SwinT introduces shifts of $(P/2, M/2, M/2)$ patches along the depth, height, and width dimensions, enhancing inter-window information interaction.

The 3D SwinT architecture comprises four stages. Each stage includes a patch merging module and multiple 3D SwinT blocks (except Stage 1). The patch merging module aggregates neighboring $2 \times 2 \times 2$ patches into larger patches, effectively reducing the spatial resolution to a quarter of its original size. A linear layer then projects the concatenated feature dimensions to half their size. The 3D SwinT blocks in each stage extract self-attention features while preserving the input resolution. Consequently, the feature map sizes at different stages are $(H/4) \times (W/4) \times (D/4) \times C$ (Stage 1), $(H/8) \times (W/8) \times (D/8) \times 2C$ (Stage 2), and so forth. Compared to standard SwinT blocks, 3D SwinT employs 3D window-based multi-head self-attention (3D W-MSA) to capture both spatial and volumetric information. Other architectural components, such as the multilayer perceptron (MLP), layer normalization (LN), and residual connections, remain unchanged from SwinT. Figure 7b depicts two adjacent 3D SwinT blocks within each stage, which can be represented by following the equation:

$$\begin{cases} \hat{y}^k = \text{3D W-MSA}(LN(y^{(k-1)})) + y^{(k-1)} \\ y^k = \text{MLP}(LN(\hat{y}^k)) + \hat{y}^k \\ \hat{y}^{k+1} = \text{3D SW-MSA}(LN(y^k)) + y^k \\ y^{k+1} = \text{MLP}(LN(\hat{y}^{k+1})) + \hat{y}^{k+1} \end{cases} \tag{13}$$

where 3D W-MSA and 3D SW-MSA represent the 3D window-based and shifted W-MSA mechanisms, respectively, and $\hat{y}^k$ and y^k are the outputs of 3D (S)W-MSA and MLP in block K, respectively.

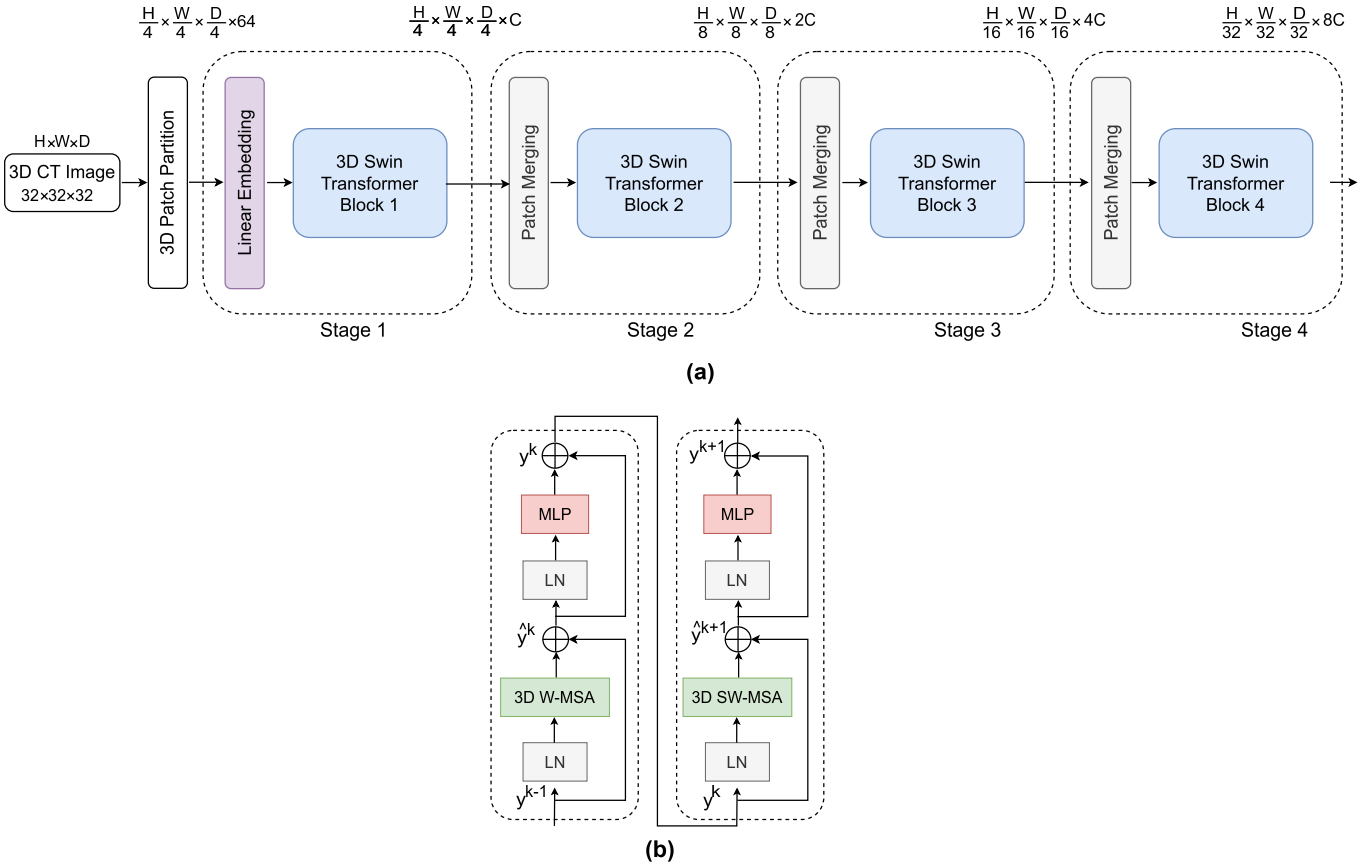

Figure 7. Overall architecture of 3DSwinT: (**a**) network architecture; (**b**) two consecutive 3DSwinT blocks.

3.5. Evaluation Metrics

To comprehensively assess the performance of the proposed model, two key evaluation metrics—recall (sensitivity) and competition performance metric (CPM)—were employed. These metrics are widely utilized in the field of computer-aided detection (CAD) to evaluate the accuracy and robustness of nodule detection systems.

Recall, or sensitivity, quantifies the model's ability to correctly identify all existing nodules within the annotated dataset. Specifically, it measures the proportion of true nodules (ground truth) that are successfully detected by the model. This metric is especially critical in medical imaging applications, where missing even a single malignant nodule can delay diagnosis and significantly affect patient outcomes. In the context of lung cancer screening, a high recall is imperative to minimize false negatives and ensure that potential cancerous regions are not overlooked.

The recall metric is mathematically defined as:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{14}$$

where *TP* denotes the number of correctly detected nodules, and *FN* represents the number of nodules present in the dataset but missed by the model. A higher recall value indicates stronger detection sensitivity and reduced clinical risk, which is particularly important in early-stage cancer detection when nodules are small and harder to detect.

The competition performance metric (CPM) measures the average sensitivity of the model across a range of false-positive rates (typically 1/8, 1/4, 1/2, 1, 2, 4, and 8 false

positives per scan). The CPM provides a holistic assessment of the model's performance, balancing its sensitivity and specificity at varying levels of false positives.

$$\text{CPM} = \frac{\sum_{i=1}^{n} \text{Sensitivity at } FP_i}{n} \tag{15}$$

where n is the number of predefined false-positive thresholds (FP_i).

4. Experimental Results and Discussion

In this section, we present a comprehensive evaluation of the proposed DA OMS-CNN framework. The results are structured into three key subsections: (1) implementation details and training setup, (2) an ablation study to assess the individual contributions of each proposed module, and (3) experimental comparisons with state-of-the-art lung nodule detection methods on both the LUNA16 and PN9 datasets. These analyses collectively demonstrate the effectiveness and generalization capabilities of our proposed approach.

4.1. Implementation

This study approaches lung nodule detection through three key stages: region proposal extraction, classification, and false-positive reduction. Initially, the DA OMS-CNN architecture is utilized for feature extraction, while the RPN is employed for training. The hyperparameters $[N_S, K_S, M_S]$ for small nodules and $[N_L, K_L, M_L]$ for large nodules are tuned prior to training, using two distinct RPNs for each category. After optimization, the values for small nodules are found to be $N_S = 8$, $K_S = 505$, and $M_S = 14$, while for large nodules, the values converge to $N_L = 3$, $K_L = 512$, and $M_L = 16$. A 10-fold cross-validation strategy is implemented to evaluate the system's performance, with stochastic gradient descent (SGD) optimization applied using a momentum factor of 0.9. Additionally, a weight decay of 0.00001 is incorporated, and the base learning rate is set at 0.0001. The training process is conducted in a computing environment equipped with two V100 GPUs and 192 GB of memory.

In the classification stage, addressing class imbalance is a crucial aspect of the classification network. This challenge is tackled by ensuring an equal distribution of positive and negative patches, utilizing the output from the trained RPN. In this method, region proposals with an intersection over union (IoU) greater than 0.7, along with the ground truth, are designated as positive patches, while an equal number of randomly selected proposals with an IoU below 0.1 are considered negative patches. This approach not only balances the classes but also increases the number of positive samples. During training, key hyperparameters were set, including an initial learning rate of 0.01 and a maximum of 150 epochs. To prevent overfitting, a weight decay of 1×10^{-4} was applied. The learning rate was adjusted at specific checkpoints: it was reduced to 0.001 after 50% of the epochs, further decreased to 0.0001 after 75%, and finally set to 0.00001 after 90% of the epochs. These modifications contributed to a more effective training process. Additionally, stochastic gradient descent (SGD) with a momentum of 0.9 was employed to enhance model performance.

In the false-positive reduction (FPR) phase, as mentioned in the classification section, nodule and non-nodule patches are first generated, with a size of $32 \times 32 \times 32$, and then augmented using patch augmentation techniques. The architectural hyperparameters for all three 3D SwinT models are set as follows: $C = 96$, and the layer configurations are 2, 2, 6, and 2 . Furthermore, the number of multi-head self-attention heads per stage is set to 3, 6, 12, and 24, respectively. The models are trained for 100 epochs using a fivefold cross-validation approach to assess performance. All three models utilize the AdamW optimizer, with the learning rate, momentum, batch size, and weight decay values set to 0.001, 0.6, 16, and 1×10^{-5} , respectively. A warm-up cosine annealing learning rate schedule is applied,

with the warm-up phase lasting for 30 steps. Figure 8 compares pulmonary nodules as detected by the proposed network against their correspondent ground-truth locations.

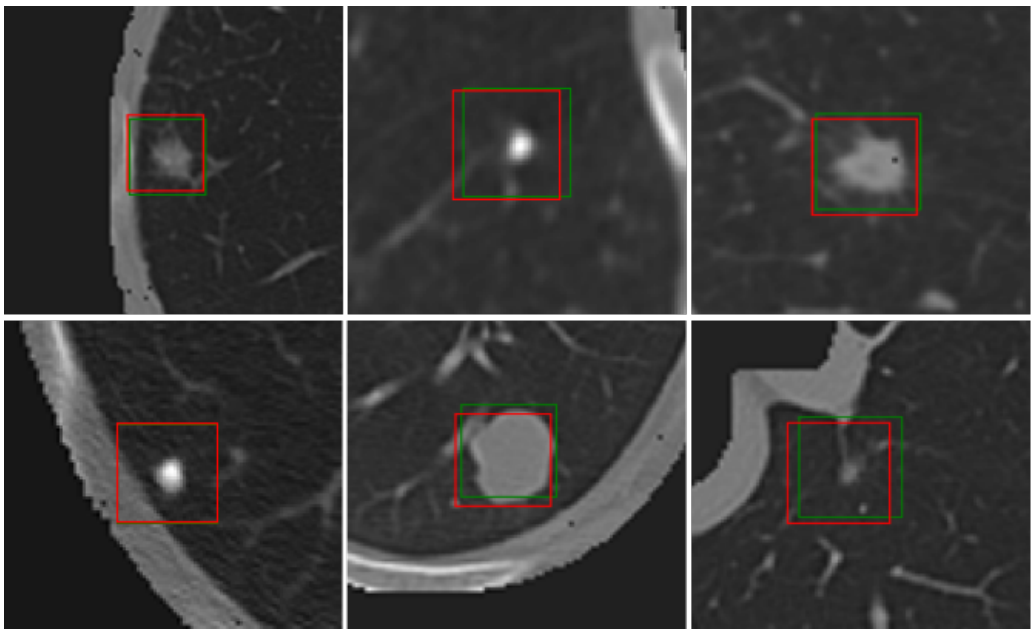

Figure 8. Pulmonary nodules detected by DA OMS-CNN (red) and their correspondent ground-truth boxes (green).

The entire model was implemented using the PyTorch deep learning framework (version 1.13). All experiments were conducted on a server equipped with two NVIDIA V100 GPUs and 192 GB of RAM.

4.2. Ablation Study

To evaluate the effectiveness of the proposed model, we conducted ablation studies under identical conditions using the LUNA16 dataset with tenfold cross-validation, as depicted in Figures 9 and 10. The ablation experiments were performed on four different configurations: (1) OMS-CNN, (2) DA OMS-CNN, (3) DA OMS-CNN with DA-RoIPooling, and (4) DA OMS-CNN with DA-RoIPooling and the proposed FPR module. This analysis helps to assess the contribution of each component to the overall model performance.

In the DA OMS-CNN approach, a dual-attention mechanism is incorporated into the final layers of OMS-CNN to enhance feature representation, as illustrated in Figure 3. This enhancement enables the extraction of high-resolution, fine-grained features, which are particularly beneficial for the early detection of lung nodules. A comparative analysis between OMS-CNN and DA OMS-CNN, presented in Figure 9, demonstrates that integrating the dual-attention mechanism into the final layers of OMS-CNN increases the average recall for 1000 region proposals by 1.3%. Additionally, as shown in Figure 10, this modification improves the CPM score from 0.839 in OMS-CNN to 0.849, further highlighting its effectiveness.

In our second contribution, we refine the classification stage by replacing RoIPooling with DA-RoIPooling. This modification enhances the model's ability to capture both spatial and channel-wise dependencies, leading to more discriminative feature representations and, ultimately, improved accuracy in lung nodule detection. To assess the effectiveness of this enhancement, Figure 9 shows that the average recall of DA OMS-CNN with DA-RoIPooling is 4.2% higher than that of OMS-CNN and 3.1% higher than DA OMS-CNN. Additionally, as depicted in Figure 10, this method achieves a CPM score of 0.86, reflecting a 2.5% improvement over the OMS-CNN approach. In the final stage, we utilize an

ensemble of three 3D SwinT models to reduce false-positive nodules. As illustrated in Figure 10, the proposed method improves the CPM score by 8.5%, 7.3%, and 5.9% compared to OMS-CNN, DA OMS-CNN, and DA OMS-CNN with DA-RoIPooling, respectively.

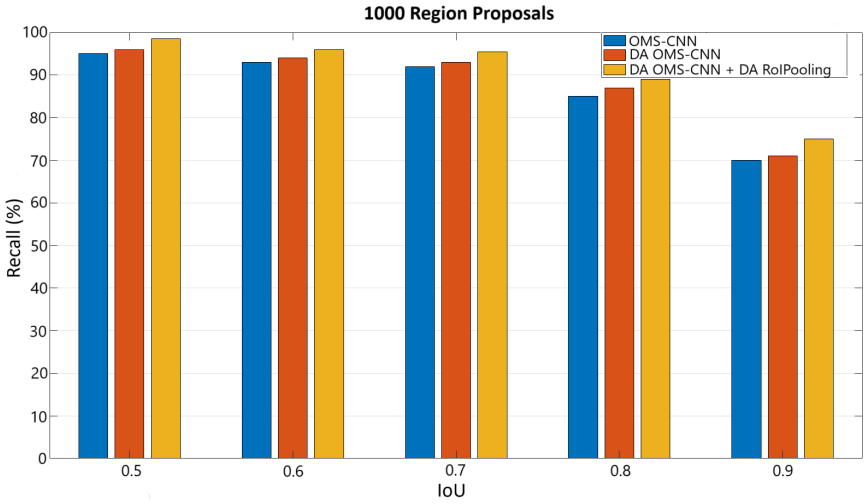

Figure 9. Recall vs. IoU overlap ration.

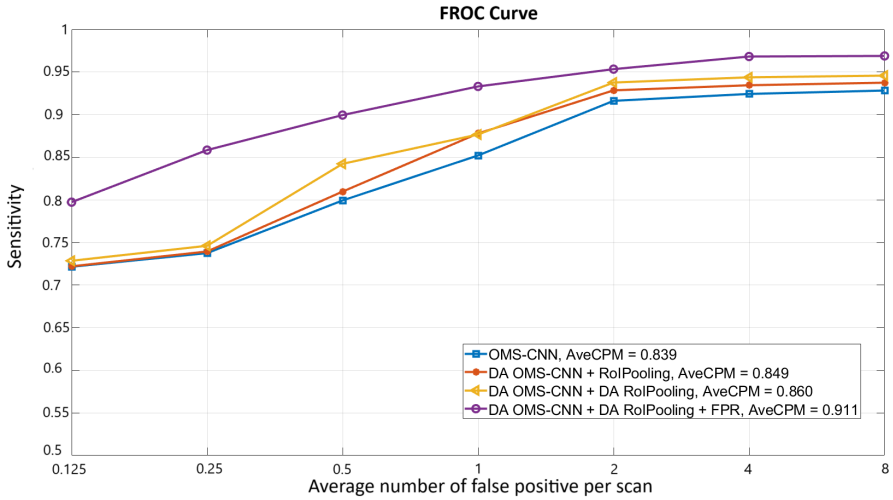

Figure 10. FROC curves of different proposed models on LUNA16.

To further clarify the impact of each proposed module, Table 1 summarizes the key results of the ablation study in a tabular format. It presents the CPM scores and sensitivity values at 1.0 false positive per scan for different configurations of our model. This complementary table enables a more intuitive comparison of performance gains achieved through the integration of dual attention mechanisms, DA-RoIPooling, and the final false-positive reduction (FPR) module. The results demonstrate the incremental improvements in both sensitivity and overall CPM, highlighting the contribution of each component to the final detection performance.

Table 1. Ablation study: performance comparison of different model configurations on LUNA16.

Model Configuration	CPM Score	Sensitivity at 1.0 FP/scan
OMS-CNN	0.839	0.8521
DA OMS-CNN	0.849	0.8967
DA OMS-CNN + DA-RoIPooling	0.860	0.9331
DA OMS-CNN + DA-RoIPooling + FPR	0.911	0.9601

4.3. Experimental Comparison

This section presents the performance evaluation of the proposed lung nodule detection framework using different experimental settings. The results are reported in three tables: Table 2 shows the performance of the proposed candidate nodule detection network before false-positive reduction on the LUNA16 dataset, Table 3 presents the results after applying false-positive reduction using the LUNA16 dataset, and Table 4 demonstrates the generalization capability of the proposed method by evaluating it on the PN9 dataset.

Table 2. Comparison of the proposed candidate nodule detection network with other methods on LUNA16.

CAD Method	Year	0.125	0.25	0.5	1.0	2.0	4.0	8.0	CPM
Dou et al. [24]	(2017)	0.6590	0.7540	0.8190	0.8650	0.9060	0.9330	0.9460	0.8390
Gu et al. [25]	(2018)	0.4801	0.6495	0.7920	0.8794	0.9163	0.9293	0.9301	0.7967
Pezeshk et al. [26]	(2018)	0.6370	0.7230	0.8040	0.8650	0.9070	0.9380	0.9520	0.8320
Xie et al. [27]	(2019)	0.4390	0.6880	0.7960	0.8520	0.8640	0.8640	0.8640	0.7750
OMS-CNN [7]	(2024)	0.7215	0.7357	0.7993	0.8521	0.9162	0.9243	0.9283	0.8396
DA OMS-CNN		0.7285	0.7461	0.8223	0.8967	0.9377	0.9438	0.9458	0.8601

Table 3. Performance comparison of different methods for false-positive reduction on LUNA16.

CAD Method	Year	0.125	0.25	0.5	1.0	2.0	4.0	8.0	CPM
Zeo et al. [28]	(2020)	0.6300	0.7530	0.8190	0.8690	0.9030	0.9150	0.9200	0.8300
CBAM [29]	(2021)	0.4670	0.6020	0.7300	0.812	0.8770	0.9150	0.9310	0.7620
I3DR-Net [30]	(2022)	0.6356	0.7131	0.7984	0.8527	0.8760	0.8992	0.9147	0.8128
MSM-CNN [23]	(2022)	0.6770	0.7410	0.8160	0.8500	0.8900	0.9050	0.9250	0.8290
MS-3DCNN [12]	(2023)	0.7280	0.7990	0.860	0.8080	0.9260	0.9410	0.9560	0.8730
AttentNet [13]	(2024)	0.7520	0.8170	0.8570	0.8850	0.9200	0.9330	0.9330	0.8710
MK-3DCNN [14]	(2024)	0.7099	0.7723	0.8356	0.8836	0.9174	0.9384	0.9562	0.8591
TED [16]	(2024)	0.7619	0.8222	0.8736	0.9069	0.9302	0.9443	0.9530	0.8846
OMS-CNN [7]	(2024)	0.7932	0.8421	0.8712	0.9048	0.9387	0.9473	0.9481	0.8922
DA OMS-CNN		0.7973	0.8584	0.8995	0.9331	0.9534	0.9682	0.9689	0.9112

Table 2 provides a comparative analysis of the proposed method against existing candidate nodule detection methods on the LUNA16 dataset. The comparison is based on the competition performance metric (CPM) score at different sensitivity thresholds. The results indicate that the proposed DA OMS-CNN method achieves the highest CPM score of 0.8601, outperforming the baseline OMS-CNN, which achieves a CPM of 0.8396. Compared to other state-of-the-art methods, such as Dou et al. [24] and Gu et al. [25], the proposed method consistently achieves higher detection performance across all sensitivity thresholds. This improvement can be attributed to the integration of domain adaptation and optimized feature extraction techniques, as well as the addition of a dual-attention mechanism in the last layers of OMS-CNN [7], which enhances the network's ability to identify candidate

nodules more effectively. Moreover, our model shows notable improvements particularly at mid-to-low false-positive rates (0.5–2.0 FP/scan), which are critical operating points in clinical screening scenarios. For example, at 1.0 FP/scan, the DA OMS-CNN achieves a sensitivity of 0.8967, significantly higher than the 0.8650 reported by Dou et al. [24] and the 0.8521 of the baseline OMS-CNN. This enhanced detection capability is mainly due to the effective integration of the dual-attention mechanism and domain adaptation strategies, which allow the model to better focus on relevant features and reduce noise from surrounding anatomical structures. These improvements contribute to the overall 2.1% increase in CPM score compared to OMS-CNN, demonstrating the practical benefits of the proposed enhancements.

Table 4. The sensitivity and CPM score compared with other methods on PN9.

CAD Method	Year	0.125	0.25	0.5	1.0	2.0	4.0	8.0	CPM
SSD512 [31]	(2016)	0.0462	0.0848	0.1476	0.2506	0.4032	0.5727	0.7080	0.3161
RetinaNet [32]	(2017)	0.0260	0.0556	0.1095	0.1925	0.2929	0.4049	0.5105	0.2274
NoduleNet [33]	(2019)	0.2117	0.3023	0.4038	0.5102	0.6129	0.7070	0.7693	0.5025
SA-Net [19]	(2021)	0.2672	0.3603	0.4746	0.5699	0.6635	0.7352	0.7832	0.5506
I3DR-Net [30]	(2022)	0.1564	0.2313	0.3700	0.5154	0.6454	0.7291	0.7753	0.4890
OMS-CNN [7]	(2024)	0.2865	0.3841	0.4775	0.5907	0.6974	0.7853	0.8432	0.5807
DA OMS-CNN		0.3015	0.3952	0.4978	0.6221	0.7205	0.8241	0.8629	0.6034

Table 3 evaluates the impact of false-positive reduction using different methods on the LUNA16 dataset. The results show that the proposed DA OMS-CNN achieves the highest CPM score of 0.9112, surpassing other state-of-the-art approaches, including TED [16] (CPM = 0.8846) and MK-3DCNN [14] (CPM = 0.8591). The improvement in performance highlights the effectiveness of the false-positive reduction strategy employed in the proposed method, which incorporates an ensemble of three 3D SwinT models. This ensemble learning approach refines the detection process, effectively reducing the number of false positives while maintaining high sensitivity for true-positive nodules. The sensitivities at 0.125, 0.25, 2, and 4 FPs/scan are 0.797, 0.858, 0.953, and 0.968, respectively, surpassing those of the best-performing method presented. Furthermore, compared to the baseline OMS-CNN [7], which achieves a CPM of 0.8922, DA OMS-CNN provides a 2.1% increase in detection accuracy, further demonstrating its robustness in distinguishing true nodules from non-nodular structures. The proposed method for detecting potential nodules demonstrates a sensitivity of 96.93%. On average, there are 9.38 candidates per scan. These results underline the significant advantage of the ensemble-based false-positive reduction strategy in balancing sensitivity and specificity. Notably, the DA OMS-CNN maintains superior sensitivity even at very low false-positive rates, which is critical for clinical usability to minimize unnecessary follow-ups. The integration of the three 3D SwinT models contributes to capturing diverse contextual features, thereby effectively filtering out false positives without compromising the true-positive detection rate. This comprehensive improvement emphasizes the robustness and practicality of the proposed approach in real-world screening settings.

To evaluate the generalization capability of the proposed method, we conducted an experiment using the PN9 dataset, and the results are presented in Table 4. The performance of the proposed approach is compared with several existing methods, including SSD512 [31], RetinaNet [32], and NoduleNet [33]. The results indicate that the DA OMS-CNN model achieves a CPM score of 0.6034, outperforming the baseline OMS-CNN (0.5807) and other existing methods such as SA-Net [19] (CPM = 0.5506) and I3DR-Net [30] (CPM = 0.4890). The consistent improvement across different sensitivity thresholds suggests that the pro-

posed method generalizes well to unseen datasets, making it a promising approach for real-world clinical applications. These findings demonstrate the strong generalization ability of the DA OMS-CNN framework beyond the primary training domain, which is essential for clinical translation, where data variability is common. The steady increase in CPM and sensitivity across various false-positive rates indicates robustness to domain shifts and dataset heterogeneity. This suggests that the combined use of dual attention mechanisms and the 3D Swin Transformer architecture effectively captures invariant and discriminative features, enabling reliable detection performance even on previously unseen datasets such as PN9.

To further understand the behavior of the proposed DA OMS-CNN model, we conducted a qualitative analysis of both successful and failed detection cases. In successful cases, the model accurately identified nodules with clear boundaries, moderate size, and strong contrast from surrounding tissues. These nodules typically appeared in central lung regions with less anatomical noise. However, the model showed reduced sensitivity in detecting extremely small nodules (less than 3mm), nodules located near complex anatomical structures such as blood vessels or the pleural wall, and in scans with low image quality or artifacts. In such cases, misclassification often resulted from insufficient contrast or structural ambiguity. Figure 11 illustrates representative examples of both detected (green box) and missed (red box) nodules. As shown, successfully detected nodules tend to be well isolated and exhibit clearer margins, while missed cases often involve small or low-contrast nodules embedded within complex anatomical surroundings. This qualitative evidence supports our earlier quantitative findings and further highlights the strengths and current limitations of the proposed framework.

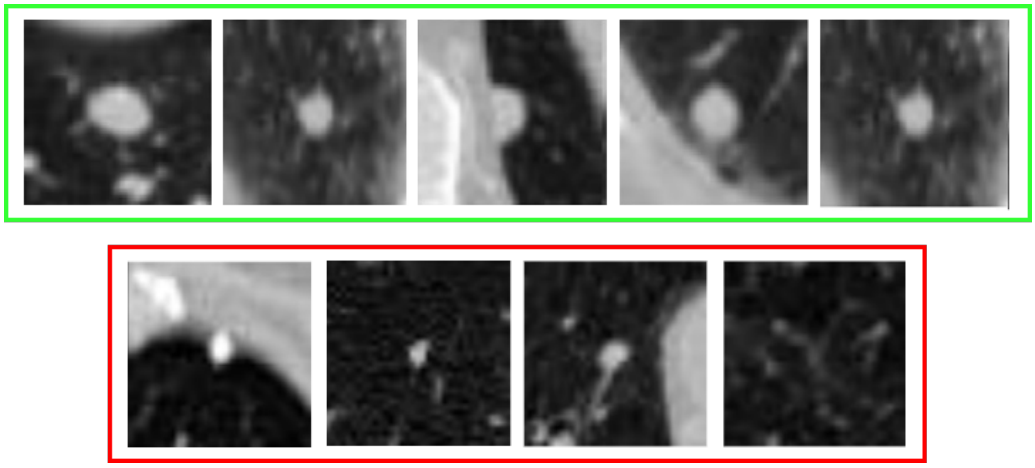

Figure 11. Examples of qualitative detection results by the proposed DA OMS-CNN. Nodules outlined in green represent correctly detected cases, while those in red indicate missed nodules.

The experimental results highlight the superior performance of the proposed DA OMS-CNN framework in lung nodule detection. The candidate nodule detection stage achieves a higher CPM score compared to existing methods, demonstrating the effectiveness of the proposed feature extraction and detection strategies. The integration of an ensemble-based false-positive reduction approach significantly enhances detection accuracy, reducing false positives while maintaining high sensitivity. Finally, the generalization experiment on the PN9 dataset further validates the robustness of the proposed method, confirming its capability to perform well on different datasets.

5. Conclusions

In this study, we presented an improved Faster R-CNN model for early-stage lung cancer detection, which integrates a novel dual-attention optimized multi-scale CNN (DA OMS-CNN) architecture and a dual-attention RoIPooling (DA-RoIPooling) technique to enhance the model's sensitivity. The DA OMS-CNN effectively captures representative features of nodules at varying sizes, while the DA-RoIPooling method further refines classification accuracy, ensuring a higher detection rate. Additionally, the incorporation of an ensemble of three 3D Swin Transformer (3D SwinT) models for false-positive reduction significantly improves the precision of the detection system. Our model demonstrated superior performance on the LUNA16 and PN9 datasets. The experimental results validate the effectiveness of the integrated DA OMS-CNN and DA-RoIPooling techniques in improving the sensitivity of lung cancer detection, while also reducing the occurrence of false-positive nodules. This advancement marks a significant step forward in the development of more accurate and reliable lung nodule detection systems, with potential applications in clinical practice.

As a future direction, we aim to enhance the clinical applicability of our system by improving its transparency and reliability. To this end, we are investigating explainable AI strategies that allow the model's decisions to be more interpretable for clinicians, helping bridge the gap between automated predictions and clinical trust. This will support the development of more user-centric and deployable CAD systems for lung cancer diagnosis.

Author Contributions: Conceptualization, Y.Z., M.R., T.O.-B. and S.M.; methodology, Y.Z. and T.O.-B.; software, Y.Z. and M.R.; validation, Y.Z.; formal analysis, Y.Z. and T.O.-B.; investigation, Y.Z. and T.O.-B.; resources, Y.Z.; data curation, Y.Z. and M.R.; writing—original draft preparation, Y.Z. and T.O.-B.; writing—review and editing, Y.Z. and T.O.-B.; visualization, Y.Z.; supervision, T.O.-B. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). The grant number is RGPIN-2021-03935.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the use of a publicly available and fully anonymized dataset (LUNA16), which does not contain any identifiable human data.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study is openly available at https://luna16.grand-challenge.org/Download/ accessed on 3 August 2023.

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

References

1. Balyan, A.K.; Ahuja, S.; Lilhore, U.K.; Sharma, S.K.; Manoharan, P.; Algarni, A.D.; Elmannai, H.; Raahemifar, K. A Hybrid Intrusion Detection Model Using EGA-PSO and Improved Random Forest Method. *Sensors* **2022**, *22*, 5986. [CrossRef]
2. Barbouchi, K.; El Hamdi, D.; Elouedi, I.; Ben Aïcha, T.; Echi, A.K.; Slim, I. A transformer-based deep neural network for detection and classification of lung cancer via PET/CT images. *Int. J. Imaging Syst. Technol.* **2023**, *33*, 1383–1395. [CrossRef]
3. Bharati, S.; Mondal, M.R.H.; Podder, P. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? *IEEE Trans. Artif. Intell.* **2023**, 5,1429–1442. [CrossRef]
4. World Health Organization. Cancer Today. Iarc.fr. Available online: https://gco.iarc.fr/today/home (accessed on 3 August 2023).
5. Dhiman, P.; Kukreja, V.; Manoharan, P.; Kaur, A.; Kamruzzaman, M.M.; Ben Dhaou, I.; Iwendi, C. A Novel Deep Learning Model for Detection of Severity Level of the Disease in Citrus Fruits. *Electronics* **2022**, *11*, 495. [CrossRef]
6. Dai, D.; Sun, Y.; Dong, C.; Yan, Q.; Li, Z.; Xu, S. Effectively fusing clinical knowledge and AI knowledge for reliable lung nodule diagnosis. *Expert Syst. Appl.* **2023**, *230*, 120634. [CrossRef]
7. Zamanidoost, Y.; Ould-Bachir, T.; Martel, S. OMS-CNN: Optimized Multi-Scale CNN for Lung Nodule Detection Based on Faster R-CNN. *IEEE J. Biomed. Health Inform.* **2024**, 29, 2148–2160. [CrossRef]

8. Jeong, Y.W.; Park, S.M.; Geem, Z.W.; Sim, K.B. Advanced parameter-setting-free harmony search algorithm. *Appl. Sci.* **2020**, *10*, 2586. [CrossRef]
9. Wu, Q.; Ma, Z.; Xu, G.; Li, S.; Chen, D. A novel neural network classifier using beetle antennae search algorithm for pattern classification. *IEEE Access* **2019**, *7*, 64686–64696. [CrossRef]
10. Gao, C.; Wu, L.; Wu, W.; Huang, Y.; Wang, X.; Sun, Z.; Xu, M.; Gao, C. Deep learning in pulmonary nodule detection and segmentation: A systematic review. *Eur. Radiol.* **2025**, *35*, 255–266. [CrossRef]
11. Zamanidoost, Y.; Alami-Chentoufi, N.; Ould-Bachir, T.; Martel, S. Efficient region proposal extraction of small lung nodules using enhanced VGG16 network model. In Proceedings of the 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, 22–24 June 2023; pp. 483–488.
12. Tan, Y.; Fu, X.; Zhu, J.; Chen, L. A improved detection method for lung nodule based on multi-scale 3D convolutional neural network. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7034. [CrossRef]
13. Almahasneh, M.; Xie, X.; Paiement, A. AttentNet: Fully Convolutional 3D Attention for Lung Nodule Detection. *arXiv* **2024**, arXiv:2407.14464. [CrossRef]
14. Wu, R.; Liang, C.; Zhang, J.; Tan, Q.; Huang, H. Multi-kernel driven 3D convolutional neural network for automated detection of lung nodules in chest CT scans. *Biomed. Opt. Express* **2024**, *15*, 1195–1218. [CrossRef] [PubMed]
15. Srivastava, D.; Srivastava, S.K.; Khan, S.B.; Singh, H.R.; Maakar, S.K.; Agarwal, A.K.; Malibari, A.A.; Albalawi, E. Early Detection of Lung Nodules Using a Revolutionized Deep Learning Model. *Diagnostics* **2023**, *13*, 3485. [CrossRef] [PubMed]
16. Ma, L.; Li, G.; Feng, X.; Fan, Q.; Liu, L. TiCNet: Transformer in Convolutional Neural Network for Pulmonary Nodule Detection on CT Images. *J. Imaging Inform. Med.* **2024**, *37*, 196–208. [CrossRef]
17. Sun, R.; Pang, Y.; Li, W. Efficient lung cancer image classification and segmentation algorithm based on an improved swin transformer. *Electronics* **2023**, *12*, 1024. [CrossRef]
18. LUNA16—Grand Challenge. Grand-Challenge.org. Available online: https://luna16.grand-challenge.org/Download/ (accessed on 3 August 2023).
19. Mei, J.; Cheng, M.M.; Xu, G.; Wan, L.R.; Zhang, H. SANet: A slice-aware network for pulmonary nodule detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4374–4387. [CrossRef]
20. Chlap, P.; Min, H.; Vandenberg, N.; Dowling, J.; Holloway, L.; Haworth A. A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* **2021**, *65*, 545–563. [CrossRef]
21. Bauckhage, C. *Numpy/Scipy Recipes for Image Processing: Binary Images and Morphological Operations*; Technical Report; B-IT, University of Bonn, Fraunhofer IAIS: Sankt Augustin, Germany, 2017.
22. UrRehman, Z.; Qiang, Y.; Wang, L.; Shi, Y.; Yang, Q.; Khattak, S.U.; Aftab, R.; Zhao, J. Effective lung nodule detection using deep CNN with dual attention mechanisms. *Sci. Rep.* **2024**, *14*, 3934 [CrossRef]
23. Zhao, Y.; Wang, Z.; Liu, X.; Chen, Q.; Li, C.; Zhao, H.; Wang, Z. Pulmonary nodule detection based on multiscale feature fusion. *Comput. Math. Methods Med.* **2022**, *2022*, 8903037. [CrossRef]
24. Dou, Q.; Chen, H.; Jin, Y.; Lin, H.; Qin, J.; Heng, P.A. Automated pulmonary nodule detection via 3d convnets with online sample filtering and hybrid-loss residual learning. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2017, Proceedings of the 20th International Conference, Quebec City, QC, Canada, 11–13 September 2017*; Proceedings, Part III 20; Springer: Berlin/Heidelberg, Germany, 2017; pp. 630–638.
25. Gu, Y.; Lu, X.; Yang, L.; Zhang, B.; Yu, D.; Zhao, Y.; Gao, L.; Wu, L.; Zhou, T. Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs. *Comput. Biol. Med.* **2018**, *103*, 220–231. [CrossRef]
26. Pezeshk, A.; Hamidian, S.; Petrick, N.; Sahiner, B. 3-D convolutional neural networks for automatic detection of pulmonary nodules in chest CT. *IEEE J. Biomed. Health Inform.* **2018**, *23*, 2080–2090. [CrossRef] [PubMed]
27. Xie, H.; Yang, D.; Sun, N.; Chen, Z.; Zhang, Y. Automated pulmonary nodule detection in CT images using deep convolutional neural networks. *Pattern Recognit.* **2019**, *85*, 109–119. [CrossRef]
28. Zuo, W.; Zhou, F.; He, Y. An embedded multi-branch 3D convolution neural network for false positive reduction in lung nodule detection. *J. Digit. Imaging* **2020**, *33*, 846–857. [CrossRef] [PubMed]
29. Sun, L.; Wang, Z.; Pu, H.; Yuan, G.; Guo, L.; Pu, T.; Peng, Z. Attention-embedded complementary-stream CNN for false positive reduction in pulmonary nodule detection. *Comput. Biol. Med.* **2021**, *133*, 104357. [CrossRef]
30. Harsono, I.W.; Liawatimena, S.; Cenggoro, T.W. Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning. *J. King Saud Univ. Comput. Inf. Sci.* **2022**, *34*, 567–577. [CrossRef]
31. Wei, L. SSD: Single shot multibox detector. In *Computer Vision ECCV2016 14th European Conference Proceedings Part I, Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.

32. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
33. Tang, H.; Zhang, C.; Xie, X. Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Proceedings of the 22nd International Conference, Shenzhen, China, 13–17 October 2019*; Proceedings, Part VI 22; Springer: Berlin/Heidelberg, Germany, 2019; pp. 266–274 .