Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 3, No. 3, 2025

HierDETR: Hierarchical Multi-Head Attention-based Pulmonary Nodule Detection in Medical Imaging

Zhenhao Tong *

Huazhong University of Science and Technology, Luoyu Road 1037, Wuhan, China

Abstract: Lung cancer is one of the leading causes of morbidity and mortality worldwide, and early detection is crucial for improving patient survival rates. Medical imaging, particularly chest CT scans, plays a vital role in early lung cancer screening and pulmonary nodule detection. However, the detection of pulmonary nodules is challenging due to the variability in size, shape, density, and location of nodules, as well as their complex background, which often includes normal or benign structures such as blood vessels and bronchial walls. To address these challenges, this paper presents a novel pulmonary nodule detection framework based on Hierarchical Multi-Head Attention (HMHA) and Query-Key Cache Updating (QKCU) mechanisms. The proposed HierDETR model improves upon existing methods by reducing redundancy in attention heads and effectively utilizing multi-scale and hierarchical context information. Experimental results on the publicly available LUNA16 dataset demonstrate that the HierDETR model outperforms mainstream methods, achieving significant improvements in detection metrics such as F1 Score, Average Precision, and Average Recall. This work provides a promising approach for enhancing the robustness and accuracy of pulmonary nodule detection, with potential applications in clinical practice for early lung cancer diagnosis.

Keywords: Pulmonary Nodule Detection; Hierarchical Multi-Head Attention; Early Lung Cancer Diagnosis.

1. Introduction

Lung cancer is one of the most prevalent and deadly malignant tumors globally, posing a severe threat to human health [1]. Numerous studies have shown that early screening and diagnosis of lung cancer are crucial for prognosis and can significantly improve the five-year survival rate of patients. Medical imaging, particularly chest computed tomography (CT), has become the most important tool for current lung cancer screening and pulmonary nodule detection due to its high resolution and three-dimensional imaging capability. Pulmonary nodules, as early imaging manifestations of lung cancer, require efficient and accurate detection and analysis to enable early cancer detection [2].

However, manual detection of pulmonary nodules is a highly challenging task. Pulmonary nodules exhibit great diversity in size, shape, density, and location, and often overlap with normal or benign structures such as blood vessels, bronchi, and pleural thickening, making the background complex and prone to misdiagnosis and missed diagnoses. Additionally, when faced with large-scale screening data, prolonged image interpretation can cause visual fatigue in doctors, leading to inter-individual variability in results. To alleviate the burden on doctors and improve detection efficiency and accuracy, the development of computer-aided detection (CADe) systems has received widespread attention. In recent years, with the rapid development of deep learning, particularly the significant success of convolutional neural networks (CNN) in image recognition tasks, deep learning-based methods for pulmonary nodule detection have become the mainstream technology and have improved detection performance to some extent [3, 4].

Despite significant progress in deep learning-based methods for pulmonary nodule detection, several challenges remain. Key issues that need to be addressed include effectively handling nodules with large scale differences, accurately distinguishing real lesions from false positives in complex backgrounds with high similarity, and fully utilizing multi-scale and hierarchical contextual information within images. Attention mechanisms [5], particularly Multi-Head Attention (MHA), have been introduced into medical image analysis tasks due to their ability to adaptively capture the relationships between features. However, the standard MHA may suffer from redundancy in the learned attention heads when processing features with complex hierarchical structures or containing different types of information, meaning that different attention heads may learn similar features or relational patterns, failing to fully leverage the benefits of parallel learning to capture diverse information. This limitation restricts its potential in complex pulmonary nodule detection tasks.

Current research mainly enhances the model's focus on lesion areas by introducing attention mechanisms [6, 7], but two key bottlenecks remain: first, different attention heads in MHA tend to learn similar features, leading to wasted computational resources and reduced feature diversity; second, existing methods do not fully utilize the hierarchical semantic structure of medical images—low-level features contain detailed information such as nodule edges, while high-level features encode global distribution patterns of the lesion. An efficient hierarchical attention architecture can achieve cross-layer feature interaction while reducing redundant computation, becoming a key to improving the robustness of pulmonary nodule detection.

The Hierarchical Multi-Head Attention (HMHA) approach [8] addresses these issues by dividing attention into subspaces of different sizes, forcing each head to learn differentiated contextual features at specific scales (such as 3×3, 5×5, and global windows), thus mitigating the redundancy problem of MHA. Furthermore, the introduction of the Query-Key Cache Updating (QKCU) module enhances the interaction of information between attention heads through an intra-layer and inter-layer update scheme, enabling the attention heads to mutually learn and collaborate, further reducing redundancy and improving the discriminative power and diversity of

learned features.

To address the above challenges, this paper introduces a pulmonary nodule detection framework based on Hierarchical Multi-Head Attention (HMHA), with key innovations including:

The novel introduction of HMHA and the QKCU mechanism, applied for the first time to the task of pulmonary nodule detection in medical imaging.

Experimental validation on public pulmonary nodule datasets, showing that the proposed method outperforms existing mainstream methods in terms of detection performance.

2. Method

2.1. Formalization of Problems

Let a lung CT scan sequence be represented as a 3D tensor, where $(H \times W)$ denotes the spatial resolution of each slice and D denotes the number of slices. The pulmonary nodule detection task can be formulated as a sparse object detection problem, with the objective function defined as:

$$f: \mathcal{J} \to \mathcal{S} = \{(p_i, b_i, s_i)\}_{i=1}^{N}$$

Where $\mathrm{p_i} \in \mathbb{Z}^3$ denotes the center coordinate of the ii-th nodule, $\mathrm{b_i} \in \mathbb{R}^6$ represents the 3D bounding box parameters (center coordinates + dimensions), $\mathrm{s_i} \in [0, 1]$ is the confidence score, and N is the number of candidate nodules.

2.2. Preprocessing and Feature Enhancement

2.2.1. Lung Parenchyma Segmentation

Otsu's thresholding method is applied to binarize the CT values $\mu \in [-1000, 1000]$ HU. The optimization objective is to maximize the between-class variance between the foreground (lung region Ω_L) and the background (chest wall/mediastinum):

$$\sigma_B^2(t) = \omega_0(t)\omega_1(t)(\mu_0(t) - \mu_1(t))^2 \quad t^* = arg\max_t \sigma_B^2(t)$$

Where ω_0, ω_1 are the proportions of foreground and background pixels, and μ_0, μ_1 are the corresponding region means.

2.2.2. Contrast-Limited Adaptive Histogram Equalization (CLAHE)

Block-based histogram equalization is applied to the segmented region Ω_L, with the constraint that the histogram distribution within each local window $\mathcal{W}_k$ does not exceed the threshold T_{clip}:

$$\widehat{\mathcal{W}}_k = \mathrm{clip}(\mathcal{W}_k, T_{clip}) \otimes \mathscr{H}_{\mathrm{eq}}$$

Where $\otimes$ denotes the histogram equalization operator, and $\mathscr{H}_{\mathrm{eq}}$ is the cumulative distribution function mapping.

2.2.3. Maximum Intensity Projection (MIP)

Adjacent d slices are merged along the axial direction to generate a 2D projection map $\mathcal{P} \in \mathbb{R}^{H \times W}$:

$$\mathcal{P}(x, y) = \max_{z \in [z_0, z_0 + d]} \mathcal{J}(x, y, z)$$

This operation enhances the cumulative intensity features of nodules across consecutive slices.

2.3. HierDETR Architecture

The model is based on the Transformer architecture, introducing Hierarchical Multi-Head Attention (HMHA) and the QKCU interaction mechanism, mathematically defined as follows:

2.3.1. Hierarchical Multi-Head Attention (HMHA)

Let the input feature $\mathbf{X} \in \mathbb{R}^{n \times d}$ be divided into K subspaces at different scales $\{\mathbf{X}_k \in \mathbb{R}^{n_k \times d_k}\}_{k=1}^{K}$, each corresponding to an attention head h_k. For the k-th head:

$$\mathbf{Q}_k = \mathbf{X}_k \mathbf{W}_k^Q, \qquad \mathbf{K}_k = \mathbf{X}_k \mathbf{W}_k^K, \qquad \mathbf{V}_k = \mathbf{X}_k \mathbf{W}_k^V$$

Where $\mathbf{W}_k^Q, \mathbf{W}_k^K, \mathbf{W}_k^V \in \mathbb{R}^{d_k \times d_h}$ are learnable parameters, and the subspaces satisfy $\sum_{k=1}^{K} n_k = n$ and $\bigcup_{k=1}^{K} \mathbf{X}_k = \mathbf{X}$. The attention weight is calculated as:

$$\mathbf{A}_k = \mathrm{Softmax}\left(\frac{\mathbf{Q}_k \mathbf{K}_k^T}{\sqrt{d_h}} \otimes \mathbf{M}_k\right)$$

Where $\mathbf{M}_k$ is the mask matrix defining the subspace neighborhood relationship (e.g., 3×3 local window or global attention).

2.3.2. QKCU Interaction Mechanism

Intra-layer Interaction: Dynamic fusion of multi-head outputs through channel attention gating $\mathbf{G}_c \in \mathbb{R}^{d_h}$:

$$\mathbf{G}_c = \sigma\big(\mathbf{W}_g \cdot \mathrm{AvgPool}(\mathbf{V}_1 \oplus \cdots \oplus \mathbf{V}_K)\big)$$

Where $\oplus$ denotes concatenation, and σ is the Sigmoid function. The final output is:

$$\mathbf{Z} = \sum_{k=1}^{K} \mathbf{G}_c^{(k)} \cdot (\mathbf{A}_k \mathbf{V}_k)$$

Inter-layer Interaction: Low-level detail features $\mathbf{F}_l$ and high-level semantic features $\mathbf{F}_h$ are passed through cross-layer skip connections:

$$\mathbf{F}_{\mathrm{fusion}} = \mathrm{LN}\big(\mathbf{F}_h + \mathbf{W}_p \cdot \mathrm{UpSample}(\mathbf{F}_l)\big)$$

Where LN denotes layer normalization, and $\mathbf{W}_p$ is the projection matrix.

2.4. Hybrid Loss Function

The total loss function is a weighted combination of Focal Loss and DETR matching loss:

$$\mathscr{L}_{\mathrm{total}} = \lambda_1 \mathscr{L}_{\mathrm{Focal}} + \lambda_2 \mathscr{L}_{\mathrm{Match}}$$

Focal Loss: Addresses the foreground-background class imbalance [9]:

$$\mathscr{L}_{\mathrm{Focal}} = -\sum_{i=1}^{N} \alpha\,(1 - p_i)^{\gamma} \log(p_i)$$

Where α is the class weight and γ is the modulation factor.

DETR Matching Loss: Optimizes detection box assignments using the Hungarian algorithm:

$$\mathscr{L}_{\mathrm{Match}} = \sum_{i=1}^{N} \left[\|\hat{b}_i - b_{\sigma(i)}\|_1 + \mathscr{L}_{\mathrm{GIoU}}(\hat{b}_i, b_{\sigma(i)})\right]$$

Where σ is the optimal assignment mapping, and $\mathscr{L}_{\mathrm{GIoU}}$ is the generalized intersection over union loss.

2.5. Experimental Setup and Optimization

Dataset: The LUNA16 dataset contains 888 CT scans, with nodule annotations following the Fleischner Society

guidelines.

Optimizer: The AdamW optimizer is used with an initial learning rate $\eta = 10^{-4}$ and cosine annealing schedule.

3. Result

To validate the effectiveness of the proposed lung nodule detection method based on Hierarchical Multi-Head Attention (HierDETR), we conducted a rigorous evaluation on the publicly available LUNA16 dataset. Comparative experiments were performed against several representative and widely used detection models, including CNN-based Yolo series models, the standard Transformer-based DETR, and Lung-DETR, which is specifically optimized for pulmonary medical imaging. The evaluation metrics included F1 Score, Average Precision (AP), and Average Recall (AR), which are commonly used in medical image detection tasks. The results are summarized in Table 1.

Table 1. Comparison of our model with SOTA on three indicators

	F1 Score	Average Precision	Average Recall
Yolo [10]	89.6	87.9	91.7
DETR [11]	91.2	90.3	93.4
Lung-DETR [12]	93.8	92.4	94.8
ours	95.1	94.6	96.7

As shown in Table 1, the proposed HierDETR achieved the best performance across all evaluation metrics. Compared with the baseline models Yolo [9], DETR [10], and Lung-DETR [11], HierDETR improved the F1 Score by 5.5%, 3.9%, and 1.3%, respectively; the Average Precision by 6.7%, 4.3%, and 2.2%; and the Average Recall by 5.0%, 3.3%, and 1.9%, respectively. Notably, even when compared with Lung-DETR, which is tailored for pulmonary imaging, HierDETR still demonstrated significant performance gains, achieving an F1 Score exceeding 95%, an Average Precision of 94.6%, and an Average Recall of 96.7%. These results strongly support the superiority and effectiveness of the proposed Hierarchical Multi-Head Attention (HMHA) and the accompanying QKCU mechanism in the task of pulmonary nodule detection.

4. Conclusion

This paper addresses the challenges in medical image-based pulmonary nodule detection, particularly the redundancy issues of the standard multi-head attention mechanism and the underutilization of hierarchical semantic information in images. We propose the HierDETR detection framework based on Hierarchical Multi-Head Attention (HMHA) and the QKCU mechanism. By innovatively introducing HMHA into the Transformer architecture, we divide the attention calculation into subspaces at different scales, forcing the attention heads to learn differentiated contextual features, effectively alleviating the redundancy problem in multi-head attention. Meanwhile, the QKCU mechanism enhances the synergy between attention heads and the fusion of cross-layer features through intra-layer and inter-layer interaction schemes, enabling the model to more robustly capture the feature information of pulmonary nodules at various scales and layers.

Experimental results on the publicly available LUNA16 dataset demonstrate that the proposed HierDETR model significantly outperforms existing mainstream detection methods such as Yolo, DETR, and Lung-DETR in key performance metrics, including F1 Score, Average Precision, and Average Recall. The improvements, particularly in Average Precision and Average Recall, indicate that our method not only identifies true nodules more accurately but also effectively reduces the false-negative rate, which is of great clinical significance for early lung cancer screening.

In summary, this paper successfully applies Hierarchical Multi-Head Attention and the QKCU mechanism to the task of medical image-based pulmonary nodule detection, and experimentally validates its effectiveness. This work provides new ideas and technical support for building more efficient and accurate computer-aided pulmonary nodule detection systems and is expected to play an important role in future clinical applications.

Discussion

The proposed HierDETR model achieves promising performance improvements in pulmonary nodule detection, primarily due to its core Hierarchical Multi-Head Attention (HMHA) and QKCU interaction mechanisms. These mechanisms effectively address two major challenges faced by standard multi-head attention when processing complex medical image features: redundancy in the learned attention head contents and insufficient utilization of hierarchical semantic information.

First, HMHA divides the attention space into subspaces at different scales (e.g., 3x3 and 5x5 local windows and global windows), forcing different attention heads to learn feature relationships at different receptive fields. This prevents all attention heads from focusing on similar global or local patterns, thereby enhancing the diversity and discriminative power of feature representations. For targets like pulmonary nodules, which exhibit large-scale variation, attention heads at different scales can focus on fine local textures of small nodules and overall shapes of larger nodules, offering an advantage over single-scale attention.

Second, the QKCU mechanism plays a key role in enhancing feature interaction. Intra-layer interaction uses channel attention gating to dynamically fuse the differentiated features learned by different attention heads, ensuring these features complement each other rather than simply overlapping. Inter-layer interaction, utilizing skip connections and upsampling, effectively combines rich spatial detail information from low-level features (e.g., nodule edges, vessel crossings) with abstract semantic information from high-level features (e.g., overall shape, relative position to surrounding tissue). This cross-layer feature fusion is crucial for accurately distinguishing true nodules from false positives in highly similar backgrounds.

Compared to baseline methods, Yolo, a typical CNN-based approach, has limited receptive fields due to the kernel size, making it less effective at capturing long-range dependencies, particularly for multi-scale nodules in complex backgrounds. Although the standard DETR introduces global attention via Transformer, its attention mechanism may be redundant, and it does not fully exploit the hierarchical structure inherent in medical images. While Lung-DETR may have adapted to pulmonary imaging through data preprocessing or model structure adjustments, the HMHA and QKCU mechanisms introduced in this paper focus on optimizing the structure and interaction of attention itself, which experimental results confirm as the key to further performance improvements.

There are some limitations to this study. First, the experiments were conducted primarily on the LUNA16

dataset, which is publicly available and relatively standardized. However, in real-world clinical applications, CT images may vary in equipment, protocols, and artifacts, so the model's generalizability needs to be validated on larger, more diverse datasets. Second, the computational complexity of Transformer models is relatively high. Although HMHA reduces computational load through hierarchical and localized approaches, further optimization is needed for real-time applications with strict performance requirements. Lastly, this study focuses on the "detection" of pulmonary nodules, which involves identifying and locating suspicious lesions, while a complete CADe system also requires benign/malignant classification of the nodules. This is a potential direction for future work.

Future research can explore the following areas: developing more efficient HMHA subspace division strategies and QKCU interaction methods to further enhance model performance and computational efficiency; extending this method to 3D CT data to better leverage complete three-dimensional nodule information; validating the model on larger, real-world datasets from different clinical centers to assess its generalization ability and clinical practicality; and integrating detection with subsequent nodule feature extraction and benign/malignant classification modules to build a more comprehensive early lung cancer diagnostic CAD system.

References

[1] Bray, Freddie, et al. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 68.6 (2018): 394-424.

[2] Javed, Rabia, et al. "Deep learning for lungs cancer detection: a review." Artificial Intelligence Review 57.8 (2024): 197.

[3] El-Bana, S., A. Al-Kabbany, and M. Sharkas. "A Two-Stage Framework for Automated Malignant Pulmonary Nodule Detection in CT Scans. Diagnostics. 2020; 10: 131."

[4] Xiao, Zhitao, et al. "Segmentation of lung nodules using improved 3D-UNet neural network." Symmetry 12.11 (2020): 1787.

[5] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[6] Walter, Joan E., et al. "Occurrence and lung cancer probability of new solid nodules at incidence screening with low-dose CT: analysis of data from the randomised, controlled NELSON trial." The Lancet Oncology 17.7 (2016): 907-916.

[7] Gu, Yu, et al. "Automatic lung nodule detection using a 3D deep convolutional neural network combined with a multi-scale prediction strategy in chest CTs." Computers in biology and medicine 103 (2018): 220-231.

[8] Zhou, Shihao, et al. "Devil is in the Uniformity: Exploring Diverse Learners within Transformer for Image Restoration." arXiv preprint arXiv:2503.20174 (2025).

[9] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

[10] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[11] Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Cham: Springer International Publishing, 2020.

[12] Ramezani, Hooman, Dionne Aleman, and Daniel Létourneau. "Lung-DETR: Deformable Detection Transformer for Sparse Lung Nodule Anomaly Detection." arXiv preprint arXiv:2409.05200 (2024).