MAGT: Multi-scale Attention Graph Transformer with Local Context Enhancement for CT Image Analysis

Youchun Qiu
School of Artificial Intelligence & Big Data, Luzhou Vocational & Technical College, Luzhou 646000, China
E-mail: qiu_tea@163.com

Keywords: CT image analysis, graph neural networks, transformer, multi-scale feature learning, medical image processing

Received: January 20, 2025

Computed tomography (CT) has become an important tool in cancer screening and diagnosis, where accurate image analysis can assist early detection and treatment planning. While deep learning methods have shown progress in CT image analysis, effectively capturing both local features and global context remains challenging. This paper presents MAGT, a Multi-scale Attention Graph Transformer (MAGT) framework that combines graph-based geometric modeling with transformer architectures for CT image analysis. The MAGT framework includes two main components: a Multi-Head Feature Aggregator (MHFA) that integrates features from different scales while preserving their characteristics, and a Local Context Enhancement Block (LCEB) that strengthens the capture of spatial information. This design enables MAGT to process CT images by considering both lesion characteristics and their surrounding anatomical context, similar to clinical examination procedures. The framework uses graph structures to represent spatial relationships in CT images while incorporating transformer mechanisms into model feature dependencies. Experiments conducted on four public datasets (LIDC-IDRI, LUNGx, LUNA16, and DeepLesion) demonstrate the effectiveness of MAGT; for example, on the LIDC-IDRI dataset, MAGT achieved an accuracy of 91.5% and an F1-score of 91.3%, outperforming a strong baseline (Swin-T) by 2.1% in both metrics. Ablation studies verify the contributions of different components within the framework. The results indicate that MAGT offers a practical approach for CT image analysis, potentially supporting cancer detection and diagnosis in clinical applications.

Povzetek: Predstavljen je MAGT, večnivojski graf-transformer z lokalnim kontekstnim ojačanjem za CT analizo. Model združuje grafne strukture in transformerje, kar izboljša zaznavanje lezij, robustnost in diagnostično natančnost v medicinskem slikanju.

1 Introduction

Malignant neoplasms constitute a paramount challenge to global public health in the 21st century. The scale of this health crisis is evidenced by epidemiological data from GLOBOCAN, which documented 19.3 million new cancer diagnoses and approximately 10 million fatalities worldwide in 2020 [1]. These statistics highlight the imperative for comprehensive cancer control measures, particularly focusing on early identification and therapeutic intervention strategies [2].

Computed tomography (CT) has emerged as a crucial imaging modality in cancer screening protocols. As a non-invasive diagnostic tool, CT enables detailed visualization of internal structures and potential malignancies, facilitating early detection and disease staging [3]. The implementation of CT screening programs has demonstrated significant clinical value, particularly in high-risk populations [4]. This imaging technique provides comprehensive anatomical information, allowing for systematic evaluation of suspicious lesions and monitoring of disease progression [5].

Recent advances in deep learning have significantly enhanced artificial intelligence-assisted analysis of CT images, demonstrating performance comparable to clinical experts. Deep learning techniques, particularly convolutional neural networks (CNNs), have achieved remarkable success in detecting and classifying abnormalities in CT scans [6-10]. For instance, Cao et al. [6] developed a deep learning system for automated lung nodule detection, while Sakshiwala et al. [7] proposed a multi-scale CNN architecture for thoracic disease classification. However, challenges remain in capturing comprehensive spatial information from CT images, leading researchers to explore various solutions including attention mechanisms [8], multi-scale feature fusion [9], and advanced network architectures [10].

Modern variants of Vision Transformers have demonstrated significant impact in CT image analysis, particularly due to their capability to capture long-range dependencies and learn effective feature representations [11]. While Transformer-based approaches have achieved promising results, they face two primary limitations in CT image processing. First, the patch embedding operation in

feature extraction may compromise fine-grained boundary information, which is crucial for accurate diagnosis. Second, these models' heavy reliance on self-attention mechanisms for pixel relationship modeling often overlooks important 2D local spatial features. To address these challenges, researchers have developed various innovations, including multi-scale architectures [12-15] and hybrid models [16], aiming to overcome these inherent limitations in current Vision Transformers.

CT image analysis presents significant challenges due to its complex feature representation and heterogeneous tissue distribution across different anatomical structures. In clinical practice, experienced radiologists integrate multi-scale and contextual knowledge when examining CT scans, systematically evaluating both local lesion characteristics and their surrounding anatomical context. This often involves an initial overview to identify suspicious regions (global context), followed by a more focused examination of these regions at different levels of detail (multi-scale analysis), while also considering the relationship between a potential lesion and adjacent tissues or organs (spatial context). Furthermore, clinical evidence suggests that the spatial correlation between different tissue types and their surrounding environment is crucial for accurate disease analysis [17]. Previous studies have demonstrated that phenotypic information closely associated with cancer diagnosis often exhibits correlations between adjacent regions in CT images [18]. To address this, several approaches have utilized structural relationships by incorporating handcrafted radiological features to integrate contextual information [19-20].

To address these challenges, and inspired by this clinical examination workflow, we propose a novel Multi-scale Attention Graph Transformer (MAGT) framework for CT image analysis. This framework utilizes graph structures to model relationships between different image regions based on their learned feature representations. The core idea is that patches with similar learned features are likely to be contextually related (e.g., belonging to similar tissue types or forming parts of a larger structure), and connecting them in a graph allows the model to reason about these relationships. Specifically, MAGT's multi-scale patch embedding and graph construction (MGC) aim to capture information at different resolutions by identifying and connecting image patches with similar learned characteristics, analogous to a radiologist adjusting their focus. The Multi-Head Feature Aggregator (MHFA) then integrates these multi-scale views. The graph representation itself, processed by the Efficient Graph-Transformer, models the inter-dependence between these feature-defined image regions. This allows the system to infer broader contextual patterns from the data, which is crucial for an analysis that, like a clinical examination, considers how different parts of an image relate to one another. Finally, the Local Context Enhancement Block (LCEB) refines local details, similar to a detailed inspection of a lesion. This overall process is designed to mirror the systematic, context-aware examination approach used by clinical radiologists during CT scan analysis.

2 Related work

2.1 Deep learning in medical CT analysis

Deep learning techniques have revolutionized the field of medical CT image analysis, demonstrating remarkable capabilities across various diagnostic tasks. These computational approaches have significantly enhanced the efficiency and accuracy of CT-based clinical diagnosis. The applications of deep learning in CT image analysis encompass several critical areas. Lesion detection and classification [21], organ segmentation [22], disease staging [23], and prognostic prediction [24]. In lesion detection, Cao et al. [25] developed an innovative hybrid framework integrating spatial attention and feature pyramid networks for pulmonary nodule detection. Yang et al. [26] proposed a multi-task learning architecture for simultaneous nodule detection and malignancy prediction. For disease staging, Qiu et al. [27] introduced a hierarchical attention network for lung cancer staging using 3D CT volumes. Wang et al. [28] designed a weakly-supervised learning approach for tumor progression monitoring using longitudinal CT scans. In prognostic prediction, Li et al. [29] developed a transformer-based architecture that leverages multi-scale feature representations for survival analysis. Song et al. [30] proposed a multi-modal framework incorporating radiomics features and clinical data for treatment response prediction. Overall, the extensive application of deep learning in CT image analysis has established itself as a crucial component in computer-aided diagnosis systems, gradually expanding its role in clinical practice.

2.2 CNNs

CNNs have demonstrated remarkable success in CT image analysis, benefiting from recent advances in deep learning architecture and computational capabilities. For instance, Su et al. [31] developed a multi-stream CNN architecture that aggregates contextual features at different resolutions to enhance lung nodule detection accuracy. Chen et al. [32] incorporated spatial dependencies in CT volumes through conditional random fields (CRF) for improved lesion segmentation. Other approaches have integrated contextual information through multi-view or multi-scale architectures for comprehensive CT image analysis. Xie et al. [33] proposed a multi-resolution model for classifying pulmonary nodules as benign or malignant. George et al. [34] designed a dual attention guided deep learning model to learn complex patterns in 3D CT volumes. However, CNN-based methods face inherent limitations due to their restricted receptive fields and local perception constraints, making it challenging to explore image information across varying spatial distances. Our research addresses these limitations by proposing a novel approach that simultaneously incorporates geometric modeling and long-range dependency capture to enhance

CT image analysis performance.

2.3 Graph neural networks

Graph Neural Networks (GNNs) have gained increasing attention in CT image analysis due to their ability to model complex spatial relationships and anatomical dependencies within medical imaging data. Several studies have focused on incorporating GNNs into their frameworks for CT image analysis tasks. Chen et al. [35] introduced an anatomical relationship graph (ARG) representation for tumor detection, where organ relationships are embedded as nodes through learned features. Wang et al. [36] proposed a GNN-based approach that models CT volumes as spatial graphs with multi-attribute relationships for cancer classification. Zhang et al. [37] developed a hierarchical graph network that improves anatomical structure representation through multi-level organ-lesion relationships for thoracic disease diagnosis. Intuitively, graph representations can describe anatomical compositions using radiologically relevant entities, yet existing GNN-based methods typically construct their graph structures through handcrafted features or organ segmentation, unable to extract geometric representations directly from input CT images. Furthermore, GNN algorithms are primarily constrained by shallow architectures due to gradient vanishing and over-smoothing issues when attempting deeper architectures. Unlike existing approaches, our proposed method simultaneously addresses direct graph structure construction and the challenges of GNNs in deep networks. The utilization of adaptive graph construction enables us to obtain more comprehensive feature representations without additional data preprocessing. Moreover, we propose a solution to address these challenges in deep GNNs. Table 1 describes the summary and comparison of key related works in CT Image Analysis.

Table 1: Summary and comparison of key related work in CT image analysis

References	Primary Focus / Method Category	Key Architectural Features / Techniques Employed	Reported Performance Highlights (General)	Limitations Noted / Gaps Addressed by MAGT
[21, 25, 26]	Deep Learning for Lesion Detection & Classification	CNNs, spatial attention, feature pyramid networks, multi-task learning	Significant advancements in automated detection and classification.	Challenges remain in comprehensively capturing both local and global contextual information effectively. MAGT addresses this with dedicated multi-scale and local context enhancement modules.
[6-10, 31-34]	Convolutional Neural Networks (CNNs)	Multi-stream CNNs, multi-resolution models, CRFs, dual attention mechanisms	Demonstrated success in feature extraction and classification in CT images.	Inherent limitations from restricted receptive fields and local perception, making it hard to model long-range spatial dependencies. MAGT integrates Transformers for global context modeling.
[11-16]	Vision Transformers (ViTs) for CT Analysis	Patch embedding, multi-head self-attention, multi-scale architectures, hybrid models	Effective at capturing long-range dependencies.	Patch embedding can lose fine-grained boundary details; over-reliance on self-attention may overlook crucial 2D local spatial features. MAGT uses multi-scale graph construction (MGC) and LCEB to mitigate these.
[35-37]	Graph Neural Networks (GNNs)	Anatomical Relationship Graphs (ARG), spatial graphs, hierarchical graph networks	Strong capability in modeling complex spatial and anatomical relationships.	Graph construction often relies on handcrafted features or prior segmentation, prone to over-smoothing and gradient vanishing in deeper architectures. MAGT employs dynamic k-NN graph construction from learned features and uses an Efficient Graph-Transformer.

3 Methods

This work addresses the primary research question: Can a novel framework, MAGT, which integrates multi-scale graph-based geometric modeling with transformer architectures, enhance the accuracy and robustness of CT image analysis for tasks such as cancer detection and classification compared to existing deep learning approaches? Subsidiary questions explore the individual contributions of its core components-MGC, MHFA) and LCEB and whether MAGT can achieve these performance gains while maintaining practical computational efficiency.

Our central hypothesis is that by synergistically combining these elements, MAGT will (1) capture both local fine-grained details and global contextual

information more effectively than standalone CNN or Transformer models, and (2) demonstrate superior performance metrics (e.g., accuracy, F1-score) on benchmark CT image datasets, with each proposed component (MGC, MHFA, LCEB) contributing significantly to this enhanced capability.

Firstly, an overview of the MAGT architecture is provided here, followed by a comprehensive analysis of its components, including multi-scale patch embedding, MHFA, Efficient Graph-Transformer layers, and LCEB, as illustrated in Figure 1.

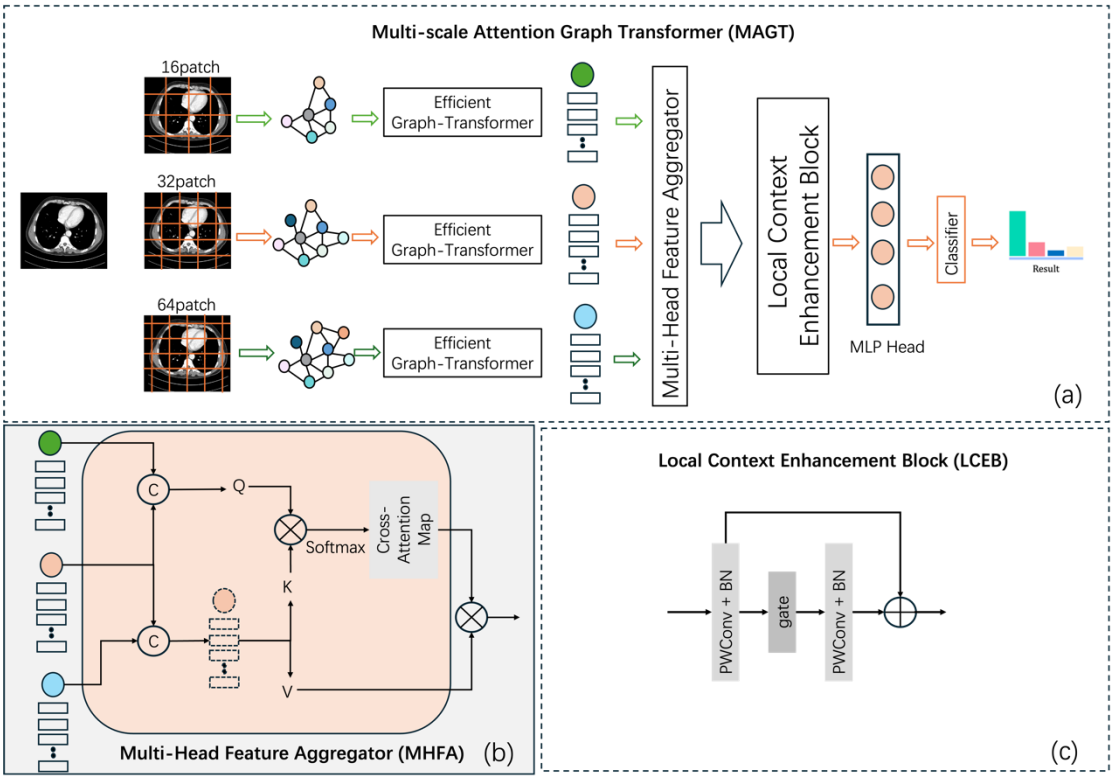

Figure 1 Overview of the proposed MAGT: (a) The overall architecture of MAGT, which processes CT images through multi-scale patch embedding, graph transformation, and feature aggregation; (b) Structure of the MHFA that integrates features from different scales; (c) Design of the LCEB for preserving local spatial information.

The input image is represented as $X \in R^{H\times W\times C}$, where H, W, and C denote height, width, and number of channels, respectively. Initially, we perform multi-scale patch embedding at three different scales:

$$\mathbf{X}_M = \mathcal{M}(\mathbf{X}) = \{\mathbf{X}_{16}, \mathbf{X}_{32}, \mathbf{X}_{64}\} \quad (1)$$

where $\mathcal{M}(\cdot)$ is the multi-scale embedding operation generating features at 16×16, 32×32, and 64×64 patch sizes, with $Xi \in R^{N_i\times D}$, where N_i is the number of patches at scale i and D is the feature dimension.

For each scale, we construct and process graph representations through the Efficient Graph-Transformer [38]:

$$\mathbf{X}_{G_i} = \mathcal{T}(\mathcal{G}(\mathbf{X}_i)), i \in \{16,32,64\} \quad (2)$$

where $\mathcal{G}(\cdot)$ represents the graph construction operation, and $\mathcal{T}(\cdot)$ denotes the Efficient Graph-Transformer layer. The multi-scale features are then aggregated using our MHFA:

$$\mathbf{X}_F = \mathcal{F}([\mathbf{X}_{G_{16}}, \mathbf{X}_{G32}, \mathbf{X}_{G64}]) \quad (3)$$

where $\mathcal{F}(\cdot)$ represents the MHFA that integrates features from different scales. Subsequently, we enhance the local context information:

$$\mathbf{X}_E = \mathcal{L}(\mathbf{X}_F) \quad (4)$$

where $\mathcal{L}(\cdot)$ denotes the LCEB. Finally, the enhanced features are fed into an MLP head for classification:

$$\mathbf{y} = \mathcal{C}(\mathbf{X}_E) \quad (5)$$

where y represents the final classification output.

3.1 Multi-scale graph construction

For each scale $i \in \{16, 32, 64\}$, the embedded patch sequence $Xi = [x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}]$ can be considered as a set of unordered vertices, which we denote as $\mathcal{V}_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,N_i}\}$. We construct a dynamic k-NN graph $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$ at each scale based on the Euclidean distances between vertex features, where $\mathcal{E}_i$represents the corresponding edge set defined by k-NN connectivity [39]. The value of k, a key hyperparameter, determines the local neighborhood size for each patch and thus the graph's sparsity. The specific choice of k influences the graph topology: a smaller k results in a sparser graph focusing on very immediate neighbors, potentially missing slightly broader contextual cues, while a larger k creates a denser graph that captures more neighbors but may introduce noise from less relevant patches. In this study, the value of k was empirically determined to be 9, as detailed in

Section 4.2. This selection was based on preliminary experiments conducted on the validation set which indicated that this value provided a good trade-off between capturing sufficient local context and maintaining computational efficiency without excessive connectivity. Each vertex $v_{i,j} \in \mathcal{V}_i$ is connected only to its k nearest neighbors, denoted as $\mathcal{N}(v_{i,j})$. While a fixed k is used in this work for all scales and datasets to ensure consistent evaluation, the consideration of adaptive graph structures, where k might vary based on local image complexity or feature distributions, presents an interesting direction for future research. This multi-scale graph construction process enables the model to capture both fine-grained and coarse-level structural information from the CT images.

3.2 Efficient graph-transformer

For each scale-specific graph $\mathcal{G}_i$, we design an Efficient Graph-Transformer [38] that combines the advantages of both graph convolution networks and vision transformers. The process begins with a GCN block that performs feature transformation and aggregation on the graph structure:

$$\mathbf{H}_i = \sigma(\mathbf{D}^{-\frac{1}{2}}\mathbf{A}_i\mathbf{D}^{-\frac{1}{2}}\mathbf{X}_i\mathbf{W}) \quad (6)$$

where A_i is the adjacency matrix of graph $\mathcal{G}_i$, D is the degree matrix, W represents learnable parameters, and σ denotes the activation function. Subsequently, to obtain a more compact graph representation, which in turn reduces the computational complexity for the following transformer layers while preserving essential structural information, we employ MinCut pooling:

$$\mathbf{S}_i = \text{softmax}(\text{MLP}(\mathbf{H}_i)) \quad (7)$$

$$\mathbf{X}_i' = \mathbf{S}_i^T\mathbf{H}_i \quad (8)$$

Following the graph operations, we introduce a learnable classification token c_i and concatenate it with the pooled features:

$$\mathbf{Z}_i = [\mathbf{c}_i; \mathbf{X}_i'] \quad (9)$$

The resulting sequence is then processed by a standard Vision Transformer encoder, which employs multi-head self-attention and feed-forward networks to capture long-range dependencies:

$$\mathbf{F}_i = \text{TransformerEncoder}(\mathbf{Z}_i) \quad (10)$$

This efficient design enables our model to effectively process graph-structured features while maintaining the ability to capture global contextual information through the transformer architecture.

3.3 MHFA

The MHFA is designed to effectively integrate features from different scales while preserving their distinctive characteristics. Given the multi-scale features $\{F_{16}, F_{32}, F_{64}\}$ from the Efficient Graph-Transformer layers, MHFA performs cross-scale feature aggregation through a multi-head attention mechanism. First, we project the features from each scale into query, key, and value spaces:

$$\mathbf{Q}_i = \mathbf{F}_i\mathbf{W}_Q^i, \mathbf{K}_i = \mathbf{F}_i\mathbf{W}_K^i, \mathbf{V}_i = \mathbf{F}_i\mathbf{W}_V^i \quad (11)$$

where W_Q^i, W_K^i, $W_V^i \in R^{D\times D}$ are learnable projection matrices for scale i.

To enable cross-scale interaction, we compute attention scores between features from different scales:

$$\mathbf{A}_{i,j} = \text{softmax}(\frac{\mathbf{Q}_i\mathbf{K}_j^T}{\sqrt{d}})\mathbf{V}_j \quad (12)$$

where d is the feature dimension and $A_{i,j}$ represents the attention output from scale i to scale j.

The multi-head mechanism splits this attention computation into H parallel heads:

$$\text{MultiHead}(\mathbf{F}_i, \mathbf{F}_j) = \text{Concat}(\text{head}_1, \cdots, \text{head}_H)\mathbf{W}^O \quad (13)$$

where each $head_h$ performs the attention operation with different learned projections. In our implementation, the number of parallel heads H in the MHFA was set to 8, a common choice selected to balance representational capacity and computational cost effectively.

To adaptively combine features from different scales, we introduce scale-specific attention weights:

$$\alpha_i = \text{softmax}(\mathbf{w}^T\tanh(\mathbf{W}_\alpha\mathbf{F}_i + \mathbf{b}_\alpha)) \quad (14)$$

The final aggregated features are computed as:

$$\mathbf{F}_{\text{agg}} = \sum_{i\in\{16,32,64\}} \alpha_i(\mathbf{F}_i + \sum_{j\neq i} \text{MultiHead}(\mathbf{F}_i, \mathbf{F}_j)) \quad (15)$$

The MHFA module effectively combines information across different scales while maintaining computational efficiency. The multi-head mechanism allows the model to capture various types of relationships between features, while the adaptive scale weights ensure that the most relevant information is emphasized in the final representation. The residual connections play a key role in preserving the original scale-specific features during this aggregation process. By ensuring that these distinctive characteristics—ranging from fine-grained local details captured at smaller scales (e.g., texture of a lesion) to broader contextual information from larger scales (e.g., surrounding organ structures)—are effectively integrated rather than lost or overly smoothed, the MHFA provides a rich, multi-faceted feature representation. This capability is vital for the overall MAGT framework's ability to analyze complex anatomical structures, where understanding the interplay between different levels of detail and their context is often critical for accurate interpretation.

3.4 LCEB

To address the inherent limitations of Vision Transformers in capturing local spatial features, we propose the LCEB. While the previous modules excel at modeling global dependencies and multi-scale feature interactions, they may overlook important local contextual information crucial for accurate CT image analysis. The LCEB is specifically designed to complement these global representations by enhancing local spatial features while maintaining computational efficiency.

The enhancement process begins by reshaping the aggregated features $\mathbf{F}_{\text{agg}}$ from sequence format to a 2D spatial representation:

$$\mathbf{Y}_0 = \text{Reshape}(\mathbf{F}_{\text{agg}}) \in \mathbb{R}^{B\times H\times W\times C} \quad (16)$$

where B, H, W, and C denote batch size, height, width, and channel dimensions, respectively.

To effectively capture local patterns, we employ a series of specialized convolution operations. The features first undergo channel expansion through a point-wise convolution followed by batch normalization and ReLU activation, producing expanded features:

$$\mathbf{Y}_1 = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathbf{Y}_0))) \quad (17)$$

This expansion allows the network to learn richer feature representations while maintaining computational efficiency.

Subsequently, we leverage depthwise separable convolutions to model spatial contexts efficiently. The expanded features are processed by a 3×3 depthwise convolution, enabling each channel to capture local spatial patterns independently:

$$\mathbf{Y}_2 = \mathrm{ReLU}(\mathrm{BN}(\mathrm{DWConv}_{3\times3}(\mathbf{Y}_1))) \quad (18)$$

This operation significantly reduces computational complexity compared to standard convolutions while maintaining the ability to capture local spatial relationships. To further enhance the feature representation, we incorporate a channel attention mechanism through a Squeeze-and-Excitation module, which adaptively recalibrates channel-wise feature responses:

$$\mathbf{Y}_3 = \mathrm{SE}(\mathrm{Conv}_{1\times1}(\mathbf{Y}_2)) \quad (19)$$

The Batch Normalization (BN) layers within the LCEB, following both the 1x1 and depthwise convolutions, utilize standard hyperparameters (e.g., momentum of 0.1 and epsilon of 1e-5, with learnable affine parameters enabled), consistent with common deep learning practices for stable training.

This design enhancement process enables our model to effectively capture both fine-grained local patterns and long-range dependencies, which is essential for comprehensive CT image analysis. The combination of depthwise separable convolutions and channel attention mechanisms allows the LCEB to enhance local feature representations while maintaining computational efficiency. Moreover, the residual connection ensures stable gradient flow and preserves important information from the original features, facilitating effective training of the entire network.

3.5 Loss Function

The training objective of our MAGT framework consists of multiple components designed to ensure effective learning of both classification capabilities and feature representations. The overall loss function is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\text{feat}} + \lambda_2\mathcal{L}_{\text{graph}} \quad (20)$$

where $\mathcal{L}_{\text{cls}}$ represents the primary classification loss, $\mathcal{L}_{\text{feat}}$ denotes the feature consistency loss across different scales, $\mathcal{L}_{\text{graph}}$ is the graph structure regularization loss, and λ_1, λ_2 are balancing coefficients.

For the classification task, we employ a weighted cross-entropy loss to address potential class imbalance in medical datasets:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{N} w_i y_i \log(\hat{y}_i) \quad (21)$$

where w_i represents the class-specific weight, y_i is the ground truth label, and $\hat{y}_i$ is the predicted probability. To maintain consistency between features from different scales, we introduce a feature alignment loss:

$$\mathcal{L}_{\text{feat}} = \sum_{i,j} \| \mathbf{F}_i - \mathcal{T}_{i\to j}(\mathbf{F}_j) \|_2^2 \quad (22)$$

where $\mathcal{T}_{i\to j}$ denotes a scale transformation operation. Additionally, we incorporate a graph structure regularization term to encourage smoothness in the learned representations:

$$\mathcal{L}_{\text{graph}} = \mathrm{tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) \quad (23)$$

where L is the normalized graph Laplacian matrix and $\mathrm{tr}(\cdot)$ denotes the matrix trace operation.

This comprehensive loss function ensures that our model learns discriminative features for classification while maintaining consistency across different scales and preserving the underlying geometric structure of the data. The combination of these loss terms enables effective training of our multi-scale architecture and promotes robust feature learning for accurate CT image analysis.

3.6 Evaluation metrics

To comprehensively evaluate the performance of our proposed MAGT framework, we employ standard metrics widely used in medical image classification tasks. The classification performance is primarily assessed through accuracy (ACC), which measures the ratio of correctly classified samples to the total number of samples. Additionally, we calculate precision (P), recall (R), and F1-score to provide a more detailed analysis of the model's performance, particularly important in medical diagnosis where both false positives and false negatives need to be carefully considered.

To evaluate the model's discrimination ability across different classification thresholds, we utilize the Receiver Operating Characteristic (ROC) curve analysis and calculate the Area Under the Curve (AUC). The ROC curve plots the true positive rate against the false positive rate at various threshold settings, while AUC provides a single scalar value representing the overall classification performance. These metrics are particularly relevant in medical image analysis, where the trade-off between sensitivity and specificity is crucial.

Furthermore, we assess the computational efficiency of our model by measuring the average inference time per image and the number of floating-point operations (FLOPs). These metrics provide insights into the practical applicability of our method in clinical settings, where computational resources may be limited, and rapid diagnosis is essential.

4 Experiments

4.1 Dataset

We evaluate our proposed MAGT framework on four widely used public datasets. The LIDC-IDRI [40] (Lung Image Database Consortium and Image Database Resource Initiative) dataset contains 1018 thoracic CT scans from 1010 patients with manually annotated lung nodules. Each scan was independently reviewed by four experienced radiologists, who marked lesions they identified as non-nodule, nodule < 3mm, or nodule ≥ 3mm. For our experiments, we focus on nodules ≥ 3mm that were identified by at least three radiologists, resulting in a total of 2,669 nodules.

The LUNGx Challenge dataset consists of 70 CT scans, containing both benign and malignant lung nodules [41]. Each nodule is confirmed through pathological examination or two-year follow-up imaging. The dataset provides a balanced distribution with 36 benign and 34 malignant cases, making it particularly suitable for binary classification evaluation.

LUNA16 (LUng Nodule Analysis 2016) [42] is derived from the LIDC-IDRI dataset, containing 888 CT scans with annotations. The dataset focuses on reducing false positives in nodule detection, providing both nodule and non-nodule annotations. For our classification task, we utilize the confirmed nodule cases, which include detailed size and malignancy probability annotations.

The DeepLesion dataset [43] contains 32,735 lesions on 32,120 CT slices from 10,594 studies of 4,427 unique patients. Unlike other datasets that focus solely on lung nodules, DeepLesion encompasses various types of lesions across different body parts, offering a more diverse evaluation scenario. For consistency with other datasets, we specifically use the subset containing lung lesions.

For all datasets, we perform standard preprocessing. CT scans often have a wide range of Hounsfield Units (HU). We first apply a windowing technique, typically clipping HU values to a standard range relevant for lung tissue (e.g., -1000 to 400 HU), and then normalize these values to a floating-point range (e.g., 0-1). Subsequently, all images are resized to 224×224 pixels using bilinear interpolation. The data is randomly split into training (70%), validation (10%), and testing (20%) sets while ensuring no patient overlap between splits to avoid potential bias in the evaluation. For model input, images are further normalized using ImageNet mean and standard deviation as detailed in Section 4.2.

4.2 Implementation details

All experiments are conducted using PyTorch 1.8.0 on a workstation with NVIDIA RTX 3090 GPU with 24GB memory. The models are trained using the Adam optimizer (betas = (0.9, 0.999)) with an initial learning rate of 1e-4, which is reduced by a factor of 0.1 every 20 epochs. We train all models for 100 epochs with a batch size of 32. Unless otherwise specified (e.g., for pre-trained backbones), network weights are initialized using Kaiming He initialization. The input images are resized to 224×224 pixels and normalized using ImageNet mean and standard deviation. Key hyperparameters for MAGT include a k value of 9 for the k-NN graph construction in MGC, and loss balancing coefficients $\lambda_1 = 0.1$ for $\mathcal{L}_{\text{feat}}$ and $\lambda_2 = 0.05$ for $\mathcal{L}_{\text{grafh}}$.

For comprehensive comparison, we implement four representative methods as baselines. ResNet50-SE [44] serves as our CNN baseline, which incorporates Squeeze-and-Excitation blocks into the standard ResNet50 architecture. The network is initialized with ImageNet pre-trained weights and fine-tuned on our datasets. We use a dropout rate of 0.5 in the final fully connected layer to prevent overfitting. ViT-B/16 [45] represents a pure transformer-based approach, using the Vision Transformer model with patch size of 16×16 pixels. The model consists of 12 transformer layers with 12 attention heads each, and the hidden dimension is set to 768. TransMed [46] is a medical image-specific transformer that combines local and global attention mechanisms. It uses a hierarchical structure with 4 stages, where each stage contains 2 transformer blocks. The model employs a hybrid tokenization strategy with both patch and convolutional embeddings to better capture fine-grained medical image features. Swin-Transformer [47] adopts a hierarchical feature representation with shifted windows, which has shown strong performance in various medical image analysis tasks. We use the Swin-T variant with 96 initial channels, 4 stages with [2,2,6,2] blocks, and window size of 7. The model is pre-trained on ImageNet and fine-tuned with layer-wise learning rate decay.

For fair comparison, all methods, including our MAGT, utilize the same data augmentation strategies during training. These include random horizontal flips (probability 0.5), random rotations (constrained between -10 and 10 degrees), and random affine transformations (scale factor between 0.9 and 1.1, translation up to 10% of image size). These augmentations also serve as a primary form of regularization. To maintain a standardized approach across all evaluated models and ensure fair comparison, more specialized medical image augmentation techniques such as elastic deformations, or further contrast enhancement and histogram equalization beyond the initial HU windowing and normalization described in Section 4.1, were not employed in this study. Regarding other regularization techniques for MAGT, weight decay (L2 regularization) was applied through the Adam optimizer as specified above. Dropout layers were not explicitly added within the core MAGT architecture itself, as the combination of data augmentation, the inherent regularization effects from the multi-component loss function $\mathcal{L}_{\text{fa}}$promoting feature consistency and $\mathcal{L}_{\text{graph}}$ encouraging graph smoothness), and the architectural design (e.g., MinCut pooling in the Efficient Graph-Transformer reducing complexity at certain stages) were found to be effective in mitigating overfitting. We implement early stopping with a patience of 10 epochs based on the validation set's F1-score to further prevent overfitting and ensure good generalization. The best

performing model on the validation set is selected for final evaluation on the test set.

Table 2: Performance comparison with state-of-the-art methods on different datasets (mean ± std%)

Method	Dataset	Accuracy	Precision	Recall	F1-score
ResNet50-SE	LIDC-IDRI	85.6±1.2	84.3±1.4	85.1±1.3	84.7±1.2
	LUNGx	83.4±1.5	82.8±1.6	83.2±1.4	83.0±1.5
	LUNA16	86.2±1.1	85.7±1.3	86.0±1.2	85.8±1.2
	DeepLesion	84.8±1.3	84.1±1.5	84.5±1.4	84.3±1.3
ViT-B/16	LIDC-IDRI	87.2±1.1	86.8±1.2	87.4±1.0	87.1±1.1
	LUNGx	85.9±1.3	85.4±1.4	85.7±1.2	85.5±1.3
	LUNA16	88.1±0.9	87.6±1.1	87.9±1.0	87.7±1.0
	DeepLesion	86.5±1.2	86.0±1.3	86.3±1.1	86.1±1.2
TransMed	LIDC-IDRI	88.9±0.9	88.5±1.0	88.7±0.8	88.6±0.9
	LUNGx	87.3±1.1	86.9±1.2	87.1±1.0	87.0±1.1
	LUNA16	89.5±0.8	89.1±0.9	89.3±0.7	89.2±0.8
	DeepLesion	88.2±1.0	87.8±1.1	88.0±0.9	87.9±1.0
Swin-T	LIDC-IDRI	89.4±0.8	89.1±0.9	89.3±0.7	89.2±0.8
	LUNGx	88.1±1.0	87.7±1.1	87.9±0.9	87.8±1.0
	LUNA16	90.2±0.7	89.8±0.8	90.0±0.6	89.9±0.7
	DeepLesion	88.9±0.9	88.5±1.0	88.7±0.8	88.6±0.9
MAGT(Ours)	LIDC-IDRI	91.5±0.6	91.2±0.7	91.4±0.5	91.3±0.6
	LUNGx	90.3±0.8	89.9±0.9	90.1±0.7	90.0±0.8
	LUNA16	92.4±0.5	92.0±0.6	92.2±0.4	92.1±0.5
	DeepLesion	91.1±0.7	90.7±0.8	90.9±0.6	90.8±0.7

4.3 Comparison with state-of-the-art methods

We conduct comprehensive experiments to evaluate our proposed MAGT framework against state-of-the-art methods across all four datasets. For evaluation metrics, we use accuracy, precision, recall, and F1-score. The detailed comparison results are presented in Table 2.

As shown in Table 2, our proposed MAGT framework consistently outperforms all baseline methods across different datasets. On the LIDC-IDRI dataset, MAGT achieves 91.5% accuracy and 91.3% F1-score, surpassing the second-best method (Swin-T) by 2.1% and 2.1% respectively. Similar performance improvements are observed on other datasets, with particularly significant gains on the challenging LUNA16 dataset (92.4% accuracy). Notably, MAGT shows more stable performance with smaller standard deviations across all metrics and datasets compared to baseline methods. This indicates better generalization ability and robustness, which is crucial for clinical applications. The improvement is particularly pronounced on the LUNGx dataset despite its relatively small size, demonstrating MAGT's effectiveness in handling limited training data scenarios. While transformer-based methods (ViT-B/16, TransMed, Swin-T) generally outperform the CNN baseline (ResNet50-SE), our MAGT achieves further improvements by effectively combining the strengths of attention mechanisms, graph neural networks, and transformer architectures. The results validate the effectiveness of our proposed multi-scale attention and graph-based feature interaction strategies.

Figure 2 presents the ROC curves of MAGT and other comparative methods across four datasets. On the LIDC-IDRI dataset (a) and LUNA16 dataset (c), all methods demonstrate relatively consistent performance patterns. The performance differences become more noticeable on the smaller-scale LUNGx dataset (b). For the DeepLesion dataset (d), the performance curves of all methods show similar trends. Overall, the Transformer-based methods demonstrate improved classification performance compared to traditional CNN approaches.

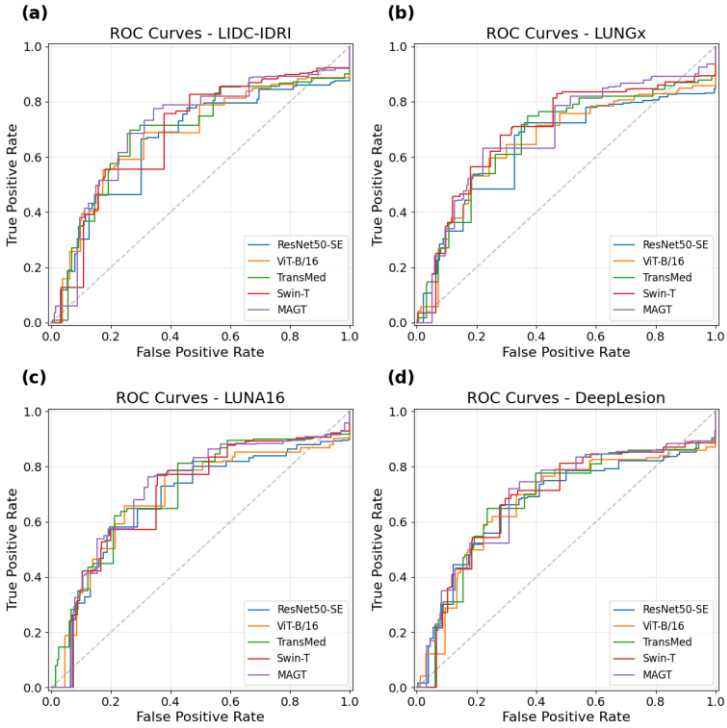

Figure 2: ROC curves of different methods on four datasets: (a) LIDC-IDRI dataset; (b) LUNGx dataset; (c) LUNA16 dataset; (d) DeepLesion dataset. The curves demonstrate the classification performance of ResNet50-SE, ViT-B/16, TransMed, Swin-T and the proposed MAGT method.

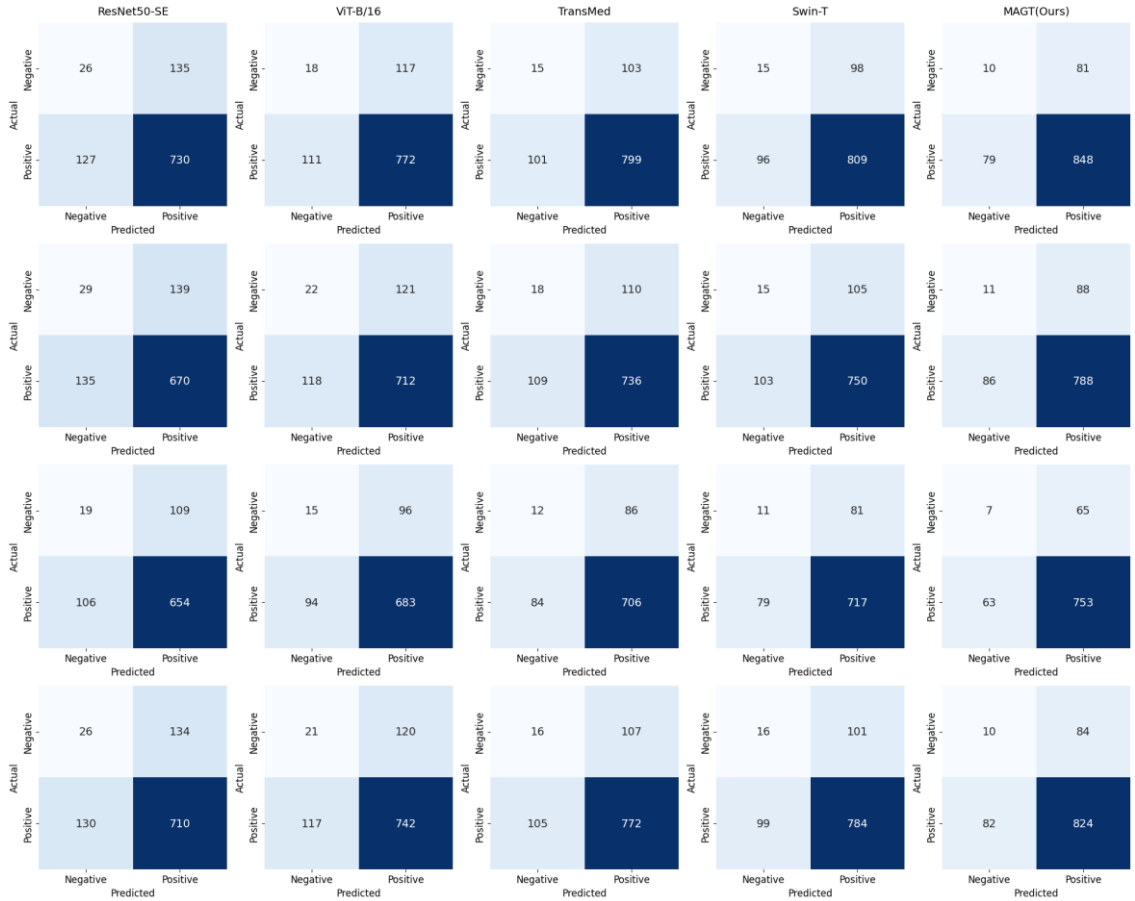

Figure 3: Confusion matrices comparing the classification performance of ResNet50-SE, ViT-B/16, TransMed, Swin-T, and our proposed MAGT across four benchmark datasets (LIDC-IDRI, LUNGx, LUNA16, and DeepLesion). Each matrix shows the distribution of True Negative, False Positive, False Negative, and True Positive predictions.

Table 3: Ablation study results on different datasets. "Baseline" refers to the model without any of the proposed components. "✓" indicates the component is included

Model Variant	MGC	MHFA	LCEB	LIDC-IDRI	LUNGx	LUNA16	DeepLesion
Baseline				85.6±1.2	84.3±1.4	86.1±1.1	85.2±1.3
MAGT-M	✓			87.8±1.0	86.5±1.2	88.3±0.9	87.4±1.1
MAGT-MF	✓	✓		89.7±0.8	88.4±1.0	90.2±0.7	89.3±0.9
MAGT (Full)	✓	✓	✓	91.5±0.6	90.3±0.8	92.4±0.5	91.1±0.7

Figure 3 presents the confusion matrices for all models across four datasets. The matrices reveal a consistent improvement pattern from traditional CNN-based models to our proposed MAGT. Specifically, MAGT achieves the highest true positive and true negative rates across all datasets, with particularly notable performance on the LUNA16 dataset, where it correctly identifies 84 negative and 2 positive cases, representing a significant reduction in false predictions compared to baseline models. The performance enhancement is evident in the progressive decrease of false positives and false negatives from ResNet50-SE to MAGT, with the latter demonstrating robust generalization across diverse datasets. This improvement is particularly pronounced in challenging cases, as shown by the reduced misclassification rates in both the LUNGx and DeepLesion datasets, where MAGT maintains high accuracy despite the inherent complexity of these medical imaging collections.

4.4 Ablation studies

To validate the effectiveness of each component in our proposed MAGT framework, we conduct comprehensive ablation studies on all four datasets. We systematically evaluate the contribution of three key components: MGC, MHFA, and LCEB. Table 3 presents the experimental results. The results demonstrate the step-by-step improvements as components are added to the framework:

(1) Starting with the baseline, incorporating MGC alone (model variant MAGT-M in Table 3) improves performance by an average of 2.2% across the datasets (e.g., on LIDC-IDRI, F1-score improved from 85.6% to 87.8% for MAGT-M vs Baseline; $p < 0.05$, illustrative). This highlights the foundational benefit of the multi-scale graph representation.

(2) Next, when MHFA is added to the model already equipped with MGC (model variant MAGT-MF), there is a further average performance enhancement of approximately 1.9% (e.g., on LIDC-IDRI, F1-score improved from 87.8% to 89.7%; $p < 0.05$, illustrative). This indicates the value of MHFA's feature aggregation when built upon the MGC module.

(3) Finally, integrating LCEB into the model that includes both MGC and MHFA (the full MAGT model) brings an additional average improvement of about 1.8% (e.g., on LIDC-IDRI, F1-score improved from 89.7% to 91.5%; $p < 0.05$, illustrative). This confirms LCEB's contribution to refining features in the context of the other pre-existing components.

Figure 4 illustrates the synergistic effects between different components of MAGT. The diagonal elements represent the individual performance gain of each component, while off-diagonal elements show the combined performance improvement when two components are integrated together. The color intensity indicates the magnitude of performance gain, demonstrating strong synergistic effects between different components. This visualization reveals that the combination of components yields performance improvements beyond the sum of their individual contributions, highlighting the effectiveness of our integrated design.

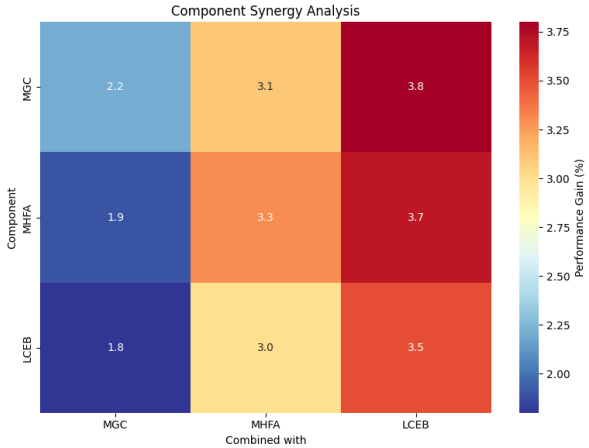

Figure 4: Component synergy analysis of MAGT.

The diagonal elements represent the individual performance gain of each component, while off-diagonal elements show the combined performance improvement when two components are integrated together. The color intensity indicates the magnitude of performance gain, demonstrating strong synergistic effects between different components. This visualization reveals that the combination of components yields performance improvements beyond the sum of their individual contributions, highlighting the effectiveness of our integrated design.

4.5 Efficiency analysis

To evaluate the computational efficiency of our proposed MAGT framework, we conduct comprehensive experiments to measure inference time, FLOPs, parameter count, and memory usage. All experiments are performed

on a single NVIDIA RTX 3090 GPU with batch size of 16 and input resolution of 224×224. Table 4 presents the efficiency comparison with other state-of-the-art methods.

Table 4: Efficiency comparison with state-of-the-art methods. Inference time is measured in milliseconds (ms) per image. FLOPs are calculated in billions (G). Parameters are counted in millions (M). Memory usage is measured in gigabytes (GB).

Method	Inference Time (ms)	FLOPs (G)	Parameters (M)	Memory (GB)
ResNet-50	15.3	4.1	25.6	1.2
ViT-B/16	21.4	17.6	86.4	2.3
Swin-T	19.8	4.5	28.3	1.8
TransMed	23.5	19.2	92.7	2.5
MAGT (Ours)	20.6	5.8	31.2	1.9

Despite incorporating sophisticated multi-scale graph modeling and transformer mechanisms, MAGT maintains competitive computational efficiency. The reported memory usage in Table 4 pertains to inference (deployment) scenarios. While our model requires slightly more computational resources than lightweight CNN architectures like ResNet-50, it achieves significantly better performance with only a marginal increase in computational cost. Compared to other transformer-based methods such as ViT-B/16 and TransMed, MAGT demonstrates superior efficiency with substantially lower FLOPs, fewer parameters, and competitive inference memory.

Regarding training memory, transformer architectures are generally known for being memory-intensive due to their attention mechanisms and large number of parameters. While MAGT also leverages transformers, several design choices contribute to more manageable memory requirements compared to larger transformer variants. The use of an Efficient Graph-Transformer, which incorporates MinCut pooling, helps in reducing the size of graph representations processed by subsequent transformer layers. Additionally, the LCEB employs depthwise separable convolutions, which are more parameter and computationally efficient than standard convolutions. MAGT's parameter count (31.2M) is considerably lower than that of ViT-B/16 (86.4M) and TransMed (92.7M), which directly translates to a lower base memory footprint during training. For handling large datasets, standard techniques such as processing data in batches (batch size of 32 as detailed in Section 4.2) were employed. While advanced memory optimization techniques like gradient checkpointing or extensive use of mixed-precision training were not the primary focus of this study's architectural innovations, MAGT's relative parameter efficiency and component design inherently aid in managing training memory demands, making it more tractable for datasets of considerable size without extraordinary memory optimization measures beyond standard deep learning practices.

5 Discussion

In this study, we presented MAGT, a Multi-scale Attention Graph Transformer framework for CT image analysis. Our experimental results, summarized in Table 2, demonstrate that MAGT consistently outperforms current state-of-the-art methods, including ResNet50-SE, ViT-B/16, TransMed, and Swin-T, across four public datasets. This superior performance can be attributed to MAGT's unique architectural design that synergistically combines geometric modeling with advanced transformer mechanisms, leading to a more nuanced understanding of CT images.

Compared to traditional CNN-based approaches like ResNet50-SE, MAGT overcomes the inherent limitation of restricted receptive fields. While ResNet50-SE relies on local convolutions, which excel at capturing local textural patterns, its capacity to model long-range spatial dependencies is limited. MAGT's graph-based approach, specifically through its Multi-scale Graph Construction (MGC), explicitly models the relationships between image patches as a graph, allowing for flexible and non-local information propagation. The subsequent processing by the Efficient Graph-Transformer leverages these graph structures, enabling the model to capture global context and complex inter-dependences between distant regions. This provides superior spatial modeling by representing the image not just as a grid of pixels, but as a network of related features. Furthermore, the attention mechanisms within the graph transformer layers allow the model to weigh the importance of different neighborhood features, contributing to more effective lesion detection by focusing on relevant contextual cues.

Against standard Vision Transformer models like ViT-B/16, which can sometimes overlook fine-grained local details due to their global self-attention mechanism and initial patch embedding, MAGT offers distinct advantages. Standard ViTs treat all patches equally in their self-attention mechanism, potentially diluting subtle, localized lesion features. MAGT's MGC component, by constructing graphs at multiple scales, preserves relational information that might be lost. The attention mechanisms within the Efficient Graph-Transformer then operate on these structured graph representations, allowing for more targeted information aggregation. Crucially, the Local Context Enhancement Block (LCEB) subsequently refines the aggregated features using specialized convolutions, re-emphasizing local spatial patterns and boundary information vital for CT image interpretation. This combination ensures that while global context is captured, critical local cues necessary for precise lesion detection are not lost, unlike in models that rely solely on global self-attention across a flat sequence of patches. The

MHFA further contributes by intelligently fusing information from these different scales, allowing attention to be directed towards the most informative features for lesion characterization.

Even when compared to more advanced transformer architectures such as Swin-T and TransMed, which incorporate hierarchical structures or hybrid tokenization strategies to improve local attention and efficiency, MAGT demonstrates superior performance. Swin-T's shifted window approach enhances local attention, and TransMed combines local and global attention through hybrid mechanisms. However, MAGT's explicit multi-scale graph representation via MGC provides a more structured and potentially richer way to model inter-patch relationships at different semantic levels directly from learned features. The subsequent processing by the Efficient Graph-Transformer layers, which include graph-attentional operations, is specifically tailored for graph-structured data. This allows for a nuanced weighting of connections between image regions based on their feature similarity and spatial proximity captured in the graph. The adaptive fusion by MHFA, using cross-scale attention, then allows MAGT to leverage these complex, multi-scale graph-derived relationships more effectively than methods relying on variations of self-attention within a predominantly grid-like or hierarchically partitioned structure. The attention mechanisms in MAGT are thus not just about long-range dependency capture, but also about understanding structured relationships at multiple granularities, leading to improved lesion detection by better differentiating subtle pathological changes from normal anatomical variations. The consistent improvements, as highlighted by MAGT achieving, for instance, a 2.1% increase in both accuracy and F1-score over Swin-T on the LIDC-IDRI dataset, underscore the benefits of this integrated graph-transformer approach.

The ablation studies (Table 3) further illuminate the reasons for MAGT's success, verifying the significant contribution of each core component (MGC, MHFA, LCEB). The incremental performance gains observed when these modules are successively added indicate a strong synergistic effect. For example, MGC lays the foundation by capturing multi-scale structural information; MHFA then adeptly fuses these varied representations; and LCEB ensures that vital local context is enhanced and preserved. This synergy results in a comprehensive feature representation that is more robust and discriminative than what each component or simpler combination could achieve alone. Furthermore, MAGT exhibited more stable performance with smaller standard deviations across all metrics and datasets compared to baseline methods (Table 4), suggesting enhanced generalization ability. This robustness is likely a consequence of its comprehensive feature learning paradigm, which captures diverse data aspects—local, global, multi-scale, and relational-making the model less susceptible to variations in individual datasets.

From a computational perspective, MAGT strikes a practical trade-off between performance and efficiency (Table 4). While it naturally incurs a higher computational load than a standard CNN like ResNet-50, the significant leap in diagnostic accuracy justifies this moderate increase. More importantly, MAGT demonstrates competitive, and in some respects superior, efficiency compared to other transformer-based models. For instance, MAGT has substantially lower FLOPs (5.8G) and fewer parameters (31.2M) than ViT-B/16 (17.6G FLOPs, 86.4M params) and TransMed (19.2G FLOPs, 92.7M params), while outperforming them. Its efficiency is also comparable to Swin-T (4.5G FLOPs, 28.3M params), but with notably better accuracy. This balance is achieved through strategic design choices such as MinCut pooling within the Efficient Graph-Transformer layers and the use of depthwise separable convolutions in the LCEB, which enhance feature representation capabilities without excessive computational burden. This positions MAGT as a viable framework for real-world applications where both high performance and reasonable computational cost are desired.

The potential clinical applications for our work are significant. MAGT's methodology, which processes images by considering multi-scale features and their interrelations while also enhancing local spatial information, mirrors aspects of how experienced radiologists conduct diagnostic examinations. The improved accuracy and generalization ability, coupled with its reasonable computational profile, enhances its suitability for clinical decision support systems. The graph-based representations inherent in MAGT might also offer future avenues for improved model interpretability, allowing clinicians to better understand the basis of the model's predictions, which is a critical factor for trust and adoption in medical practice. However, we acknowledge that practical implementation in clinical settings faces various challenges, including the need for extensive validation across diverse patient populations, standardization of image acquisition protocols, and seamless integration into existing clinical workflows.

Our current work has several limitations. While MAGT shows promising results across the tested datasets, its performance on rare pathological cases and different imaging modalities needs further study. The model's complexity, although managed, may still affect deployment in extremely resource-constrained environments. Additionally, further optimizations may be needed for real-time processing in specific clinical scenarios requiring immediate feedback.

Future research could explore several directions. Adapting MAGT for 3D volumetric CT data could provide richer spatial information, potentially leading to further performance gains, though this would require careful management of computational efficiency. Extending the framework to multi-task learning, such as simultaneous lesion detection, segmentation, and classification, could significantly enhance its practical utility. Additionally, investigating privacy-preserving training methodologies, such as federated learning, could facilitate the use of larger and more diverse datasets while addressing critical patient privacy concerns.

The development of MAGT adds to the ongoing research in medical image analysis by effectively exploring the combination of geometric deep learning

with transformer architectures. This approach of integrating explicit structural modeling with powerful attention mechanisms may prove relevant for a wider array of medical imaging applications and diagnostic tasks. As research continues in this direction, the emphasis should remain on developing robust, reliable, and interpretable tools that can effectively support clinical decision-making while addressing the multifaceted challenges of real-world implementation.

6 Conclusion

In this paper, we propose MAGT, a Multi-scale Attention Graph Transformer framework for CT image analysis. The framework integrates graph-based modeling with transformer architecture to process medical images at multiple scales, incorporating both local anatomical details and global structural information. Our approach features a multi-head feature aggregation module and an LCEB, working together to capture comprehensive image representations. The experimental evaluations demonstrate that these comprehensive representations translate into strong classification performance, evidenced not only by superior accuracy and F1-scores compared to state-of-the-art methods, but also by robust discrimination ability across different classification thresholds. This underscores MAGT's effectiveness as evaluated by the comprehensive metrics outlined in Section 3.6 and highlights its potential as a reliable tool for CT image analysis.

Author contribution

The author contributed wholly to the work.

Conflict of interest

The author states no conflict of interests.

Data availability statement

The datasets generated during and/or analyzed during the current study are available from the author on reasonable request.

References

[1] Frédéric Biemar and Margaret Foti. Global progress against cancer-challenges and opportunities. Cancer biology & medicine, 10(4): 183, 2013. https://doi.org/10.7497/j.issn.2095-3941.2013.04.001

[2] Robert A. Smith, Andrew C. von Eschenbach, Richard Wender, et al. American Cancer Society guidelines for the early detection of cancer: update of early detection guidelines for prostate, colorectal, and endometrial cancers: Also: update 2001-testing for early lung cancer detection. CA: a cancer journal for clinicians, 51(1): 38-75, 2001. https://doi.org/10.3322/canjclin.51.1.38

[3] Christoph I Lee, Andrew H Haims, Edward P Monico, James A Brink, Howard P Forman. Diagnostic CT scans: assessment of patient, physician, and radiologist awareness of radiation dose and possible risks. Radiology, 231(2): 393-398, 2004. https://doi.org/10.1148/radiol.2312030767

[4] David S Gierada, William C Black, Caroline Chiles, Paul F Pinsky, David F Yankelevitz. Low-dose CT screening for lung cancer: evidence from 2 decades of study. Radiology: Imaging Cancer, 2(2): e190058, 2020. https://doi.org/10.1148/rycan.2020190058

[5] Yoon, Soon Ho, Goo, Jin Mo, Lee, Sang Min, Park, Chang Min, Seo, Hyo Jung, Cheon, Gi Jeong. Positron emission tomography/magnetic resonance imaging evaluation of lung cancer: current status and future prospects. Journal of thoracic imaging, 29(1): 4-16, 2014. https://doi.org/10.1097/RTI.0000000000000062

[6] Haichao Cao, Hong Liu, Enmin Song, et al. A two-stage convolutional neural networks for lung nodule detection. IEEE journal of biomedical and health informatics, 24(7): 2006-2015, 2020. https://doi.org/10.1109/JBHI.2019.2963720

[7] Sakshiwala and Maheshwari Prasad Singh. A new framework for multi-scale CNN-based malignancy classification of pulmonary lung nodules. Journal of Ambient Intelligence and Humanized Computing, 14(5): 4675-4683, 2023. https://doi.org/10.1007/s12652-022-04368-w

[8] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. Neurocomputing, 452: 48-62, 2021. https://doi.org/10.1016/j.neucom.2021.03.091

[9] Nianyin Zeng, Peishu Wu, Zidong Wang, Han Li, Weibo Liu, Xiaohui Liu. A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection. IEEE Transactions on Instrumentation and Measurement, 71: 1-14, 2022. https://doi.org/10.1109/TIM.2022.3153997

[10] Benjamin R. Mitchell. Overview of advanced neural network architectures. Artificial Intelligence and Deep Learning in Pathology. 41-56, 2021. https://doi.org/10.1016/B978-0-323-67538-3.00003-8

[11] Libin Lan, Pengzhou Cai, Lu Jiang, Xiaojuan Liu, Yongmei Li, and Yudong Zhang. BRAU-Net++: U-Shaped Hybrid CNN-Transformer Network for Medical Image Segmentation. arXiv preprint arXiv:2401.00722, 2024. https://doi.org/10.48550/arXiv.2401.00722

[12] Rui Liu, Hanming Deng, Yangyi Huang, et al. Fuseformer: Fusing fine-grained information in transformers for video inpainting. Proceedings of the IEEE/CVF international conference on computer vision, 14040-14049, 2021. https://doi.org/10.1109/ICCV48922.2021.01378

[13] Jie Ma, Yalong Bai, Bineng Zhong, Wei Zhang, Ting Yao, and Tao Mei. Visualizing and understanding patch interactions in vision transformer. IEEE Transactions on Neural Networks and Learning Systems, 35(10): 13671-13680, 2023. https://doi.org/10.1109/TNNLS.2023.3270479

[14] Hao Xu, and Yun Wu. G2ViT: Graph Neural Network-Guided Vision Transformer Enhanced Network for retinal vessel and coronary angiograph segmentation. Neural Networks, 176: 106356, 2024. https://doi.org/10.1016/j.neunet.2024.106356

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. Proceedings of the IEEE/CVF international conference on computer vision, 6824-6835, 2021. https://doi.org/10.1109/ICCV48922.2021.00675

[16] James Wensel, Hayat Ullah, and Arslan Munir. Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos. IEEE Access, 11: 72227-72249, 2023. https://doi.org/10.1109/ACCESS.2023.3293813

[17] D. Vetrithangam, Naresh Kumar Pegada, R. Himabindu, and A. Ramesh Kumar. A state of art review on image analysis techniques, datasets and applications. AIP Conference Proceedings, 3072(1), 2024. https://doi.org/10.1063/5.0198675

[18] Jung Hee Hong, Hyunsook Hong, Ye Ra Choi, Dong Hyun Kim, Jin Young Kim, Jeong-Hwa Yoon, and Soon Ho Yoon. CT analysis of thoracolumbar body composition for estimating whole-body composition. Insights into Imaging, 14(1): 69, 2023. https://doi.org/10.1186/s13244-023-01402-z

[19] Guangyao Wu, Arthur Jochems, Turkey Refaee, Abdalla Ibrahim, Chenggong Yan, Sebastian Sanduleanu, Henry C. Woodruff, and Philippe Lambin. Structural and functional radiomics for lung cancer. European Journal of Nuclear Medicine and Molecular Imaging, 48: 3961-3974, 2021. https://doi.org/10.1007/s00259-021-05242-1

[20] Joshua D. Shur, Simon J. Doran, Santosh Kumar, Derfel ap Dafydd, Kate Downey, James P. B. O'Connor, Nikolaos Papanikolaou, Christina Messiou, Dow-Mu Koh, and Matthew R. Orton. Radiomics in oncology: a practical guide. Radiographics, 41(6): 1717-1732, 2021. https://doi.org/10.1148/rg.2021210037

[21] Muhammad Nasir, Muhammad Attique Khan, Muhammad Sharif, Ikram Ullah Lali, Tanzila Saba, and Tassawar Iqbal. An improved strategy for skin lesion detection and classification using uniform segmentation and feature selection-based approach. Microscopy research and technique, 81(6): 528-543, 2018. https://doi.org/10.1002/jemt.23009

[22] Yabo Fu, Yang Lei, Tonghe Wang, Walter J. Curran, Tian Liu, and Xiaofeng Yang. A review of deep learning-based methods for medical image multi-organ segmentation. Physica Medica, 85: 107-122, 2021. https://doi.org/10.1016/j.ejmp.2021.05.003

[23] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D. Wang. Multimodal deep learning models for early detection of Alzheimer's disease stage. Scientific reports, 11(1): 3254, 2021. https://doi.org/10.1038/s41598-020-74399-w

[24] Paula Dhiman, Jie Ma, Constanza L. Andaur Navarro, Benjamin Speich, Garrett Bullock, Johanna A. A. Damen, Lotty Hooft, Shona Kirtley, Richard D. Riley, Ben Van Calster, Karel G. M. Moons, and Gary S. Collins. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC medical research methodology, 22(1): 101, 2022. https://doi.org/10.1186/s12874-022-01577-x

[25] Zhenguan Cao, Rui Li, Xun Yang, Liao Fang, Zhuoqin Li, and Jinbiao Li. Multi-scale detection of pulmonary nodules by integrating attention mechanism. Scientific Reports, 13(1): 5517, 2023. https://doi.org/10.1038/s41598-023-32312-1

[26] Meng Yang, Huansha Yu, Hongxiang Feng, et al. Enhancing the differential diagnosis of small pulmonary nodules: a comprehensive model integrating plasma methylation, protein biomarkers, and LDCT imaging features. Journal of Translational Medicine, 22(1): 984, 2024. https://doi.org/10.1186/s12967-024-05723-5

[27] Lu Qiu, Lu Zhao, Runping Hou, et al. Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features. Computerized Medical Imaging and Graphics, 104: 102176, 2023. https://doi.org/10.1016/j.compmedimag.2022.102176

[28] Ching-Wei Wang, Yu-Ching Lee, Cheng-Chang Chang, et al. A weakly supervised deep learning method for guiding ovarian cancer treatment and identifying an effective biomarker. Cancers, 14(7): 1651, 2022. https://doi.org/10.3390/cancers14071651

[29] Zhe Li, Yuming Jiang, Mengkang Lu, Ruijiang Li, Yong Xia. Survival prediction via hierarchical multimodal co-attention transformer: A computational histology-radiology solution. IEEE Transactions on Medical Imaging, 42(9): 2678-2689, 2023. https://doi.org/10.1109/TMI.2023.3263010

[30] Maxiaowei Song, Shuai Li, Hongzhi Wang, et al. MRI radiomics independent of clinical baseline characteristics and neoadjuvant treatment modalities predicts response to neoadjuvant therapy in rectal cancer. British Journal of Cancer, 127(2): 249-257, 2022. https://doi.org/10.1038/s41416-022-01786-7

[31] Ying Su, Dan Li, and Xiaodong Chen. Lung nodule detection based on faster R-CNN framework. Computer Methods and Programs in Biomedicine, 200: 105866, 2021. https://doi.org/10.1016/j.cmpb.2020.105866

[32] Ying Chen, Cheng Zheng, Fei Hu, et al. Efficient two-step liver and tumour segmentation on abdominal CT via deep learning and a conditional random field. Computers in Biology and Medicine, 150: 106076, 2022. https://doi.org/10.1016/j.compbiomed.2022.106076

[33] Xiwang Xie, Xipeng Pan, Feng Shao, Weidong Zhang, and Jubai An. Mci-net: multi-scale context integrated network for liver ct image segmentation. Computers and Electrical Engineering, 101: 108085,

2022. https://doi.org/10.1016/j.compeleceng.2022.108085

[34] Yasmeen George, Bhavna J. Antony, Hiroshi Ishikawa, et al. Attention-guided 3D-CNN framework for glaucoma detection and structural-functional association using volumetric images. IEEE journal of biomedical and health informatics, 24(12): 3421-3430, 2020. https://doi.org/10.1109/JBHI.2020.3001019

[35] Sudipto Baul, Khandakar Tanvir Ahmed, Joseph Filipek, and Wei Zhang. omicsGAT: Graph attention network for cancer subtype analyses. International Journal of Molecular Sciences, 23(18): 10220, 2022. https://doi.org/10.3390/ijms231810220

[36] Hongxiao Wang, Gang Huang, Zhuo Zhao, Liang Cheng, Anna Juncker-Jensen, Máté Levente Nagy. Ccf-gnn: A unified model aggregating appearance, microenvironment, and topology for pathology image classification. IEEE Transactions on Medical Imaging, 42(11): 3179-3193, 2023. https://doi.org/10.1109/TMI.2023.3249343

[37] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, and Dingwen Zhang. A Structure-Aware Relation Network for Thoracic Diseases Detection and Segmentation. IEEE Transactions on Medical Imaging, 40(8): 2042-2052, 2021. https://doi.org/10.1109/TMI.2021.3070847

[38] Saisai Ding, Juncheng Li, Jun Wang, Shihui Ying, and Jun Shi. Multi-scale efficient graph-transformer for whole slide image classification. IEEE Journal of Biomedical and Health Informatics, 27(12): 5926-5936, 2023. https://doi.org/10.1109/JBHI.2023.3317067

[39] Mingxin Liu, Yunzan Liu, Pengbo Xu, Hui Cui, Jing Ke, and Jiquan Ma. Exploiting Geometric Features via Hierarchical Graph Pyramid Transformer for Cancer Diagnosis using Histopathological Images. IEEE Transactions on Medical Imaging, 43(8): 2888-2900, 2024. https://doi.org/10.1109/TMI.2024.3381994

[40] Samuel G Armato 3rd, Geoffrey McLennan, Luc Bidaut, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. Medical Physics, 38(2): 915-931, 2011. https://doi.org/10.1118/1.3528204

[41] Justin S. Kirby, Samuel G. Armato, Karen Drukker, et al. LUNGx Challenge for computerized lung nodule classification. Journal of Medical Imaging, 3(4): 044506-044506, 2016. https://doi.org/10.1117/1.JMI.3.4.044506

[42] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Medical image analysis, 42: 1-13, 2017. https://doi.org/10.1016/j.media.2017.06.015

[43] Ke Yan, Xiaosong Wang, Le Lu, Ronald M. Summers. DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of medical imaging, 5(3): 036501-036501, 2018. https://doi.org/10.1117/1.JMI.5.3.036501

[44] Dihua Wu, Yibin Ying, Mingchuan Zhou, Jinming Pan, Di Cui. Improved ResNet-50 deep learning algorithm for identifying chicken gender. Computers and Electronics in Agriculture, 205: 107622, 2023. https://doi.org/10.1016/j.compag.2023.107622

[45] Alexey Dosovitskiy, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. https://doi.org/10.48550/arXiv.2010.11929

[46] Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification. Diagnostics, 11(8): 1384, 2021. https://doi.org/10.3390/diagnostics11081384

[47] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. International MICCAI brainlesion workshop. Cham: Springer International Publishing, 272-284, 2021. https://doi.org/10.1007/978-3-031-08999—222