MELBA Journal Machine Learning for Biomedical Imaging

Structured Spectral Graph Representation Learning for Multi-label Abnormality Analysis from 3D CT Scans

Theo **Di Piazza** [1,2], Carole **Lazarus** [3], Olivier **Nempont** [3], Loic **Boussel** [1, 2]

1 INSA Lyon, University of Lyon, CNRS, INSERM, CREATIS UMR 5220, U1294, Villeurbanne, France
2 Hospices Civil de Lyon, Lyon, France
3 Philips Clinical Informatics, Innovation Paris, France

Abstract

With the growing volume of CT examinations, there is an increasing demand for automated tools such as organ segmentation, abnormality detection, and report generation to support radiologists in managing their clinical workload. Multi-label classification of 3D Chest CT scans remains a critical yet challenging problem due to the complex spatial relationships inherent in volumetric data and the wide variability of abnormalities. Existing methods based on 3D convolutional neural networks struggle to capture long-range dependencies, while Vision Transformers often require extensive pre-training on large-scale, domain-specific datasets to perform competitively. In this work, we propose a 2.5D alternative by introducing a new graph-based framework that represents 3D CT volumes as structured graphs, where axial slice triplets serve as nodes processed through spectral graph convolution, enabling the model to reason over inter-slice dependencies while maintaining complexity compatible with clinical deployment. Our method, trained and evaluated on 3 datasets from independent institutions, achieves strong cross-dataset generalization, and shows competitive performance compared to state-of-the-art visual encoders. We further conduct comprehensive ablation studies to evaluate the impact of various aggregation strategies, edge-weighting schemes, and graph connectivity patterns. Additionally, we demonstrate the broader applicability of our approach through transfer experiments on automated radiology report generation and abdominal CT data. This work extends our previous contribution presented at the MICCAI 2025 EMERGE Workshop. A project page is available at `https://theodpzz.github.io/projects/ctssg/`.

Keywords

3D Medical Imaging, Computed Tomography, Representation Learning, Graph Neural Network, Spectral domain, Multi-label Abnormality Classification, Automated Report Generation

Article informations

`https://doi.org/10.59275/j.melba.2026-87e3`
Volume 2026, Received: 2025-10-27, Published 2026-06-15
Corresponding author: theo.dipiazza@creatis.insa-lyon.fr

1. Introduction

Computed Tomography (CT) is a cornerstone imaging modality in clinical practice, providing radiologists with detailed three-dimensional views of the thorax and enabling the accurate detection of a wide range of abnormalities (Patel and De Jesus, 2024). However, the increasing volume of chest CT scans poses significant challenges for radiologists, who face mounting demands and time constraints (Broder and Warshauer, 2006). This has created an urgent need for automated systems capable of assisting healthcare professionals to manage their increasing workload (Najjar, 2023).

In medical imaging, early developments in automated abnormality detection predominantly focused on 2D modalities, facilitated by the availability of large-scale datasets such as CheXpert (Irvin et al., 2019) and MIMIC-CXR (Johnson et al., 2019). Early work on 3D chest CT abnormality classification initially addressed single-label classification, targeting one abnormality at a time (Panwar et al., 2020). Yet, multi-label abnormality classification is of paramount importance for clinical decision support, as it allows simultaneous detection of multiple co-occurring abnormalities and leverages inter-abnormality relationships to improve diagnostic performance (Draelos et al., 2021). Moreover, multi-label classification serves as a versatile pretext task that can later be fine-tuned for more specialized objectives,

such as report generation or disease progression modeling (Tanida et al., 2023).

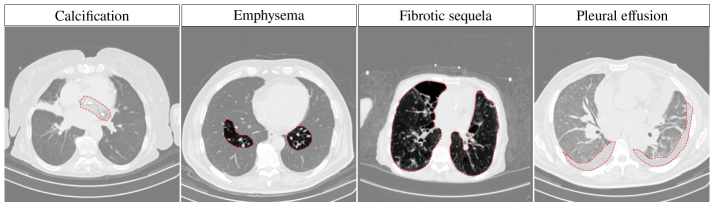

Figure 1: Axial slices from 3D CT Scans from CT-RATE, with abnormalities manually contoured in red, illustrating distinct visual characteristics.

Despite its clinical relevance, multi-label abnormality classification in 3D chest CT remains a highly challenging task due to the broad diversity of abnormalities, as illustrated in Figure 1. Furthermore, the volumetric nature of CT data necessitates the development of computationally efficient architectures that are scalable and suitable for real-world clinical deployment (Aravazhi et al., 2025).

Early approaches to multi-label abnormality classification in CT imaging predominantly leveraged fully convolutional networks. The recent release of CT-RATE, a large-scale public dataset (Hamamci et al., 2026) containing chest CT scans from over 21,000 unique patients paired with radiology reports, has significantly broadened the scope of CT-based research. This includes tasks such as synthetic volume generation (Hamamci et al., 2023) and automatic report generation (Hamamci et al., 2024). Notably, many of these methods adopt visual encoder architectures based on video vision transformers (Arnab et al., 2021), which model the 3D CT Scan as a set of 3D patches. However, 3D Transformers-based methods often rely on extensive pre-training on large, domain-specific datasets to achieve competitive performance (Hamamci et al., 2023). Recently, prior work has empiricaly demonstrated that 2.5D modeling, representing a CT volume as a set of slices rather than a set of 3D patches, can outperform purely 3D approaches. For instance, CT-Net introduced a 2.5D alternative to full 3D CNNs by modeling CT volumes as sequences of axial slices, processed independently using a 2D backbone (Draelos et al., 2021). While CT-Net demonstrated strong performance on 83 abnormalities within an internal dataset, its generalization across datasets and tasks remained unexplored, largely due to the scarcity of publicly available annotated CT datasets at the time.

Prior works describe graphs as "the main modality of data we receive from nature" (Veličković, 2023). From this perspective, most machine learning applications can be seen as special cases of graph representation learning, including Transformers (Joshi, 2025), which has lead to significant efforts in recent years across various domains of application (Zhou et al., 2020). Specifically, Transformers operate on fully connected graphs, where attention mechanisms learn adaptive edge weights between all node pairs (Giovanni et al., 2023). While this formulation has proven expressive, it entails dense connectivity (Fey and Lenssen, 2019), require extensive pre-training (Bommasani et al., 2022) and is useful for tasks where we do not have an apriori graph structure (Jumper et al., 2021), which may be suboptimal for modeling localized spatial dependencies inherent to 3D medical volumes. In contrast, representing 3D CT scans as structured graphs provides a more flexible and physically grounded framework: it allows explicit control over neighborhood definitions, edge weighting strategies, and hierarchical topologies. Recent advances in Graph Neural Networks (GNN) have demonstrated their ability to model complex relational structures across diverse imaging modalities (Ahmedt-Aristizabal et al., 2021). In medical imaging, GNNs have been successfully applied to tasks such as automated report generation (Liu et al., 2021a), where they capture semantic dependencies among knowledge entities, and whole-slide image analysis, where hierarchical graph representations enhance abnormality classification (Guo et al., 2023). These successes suggest that GNNs hold strong potential for extending 3D modeling paradigms in chest CT analysis, particularly in scenarios where volumetric context and inter-slice dependencies are critical.

Building on the representational flexibility of 2.5D approaches and the relational expressiveness of graph neural networks, we introduce CT-SSG (**S**tructural **S**pectral **G**raph for **C**omputed **T**omography), a framework that formally represents 3D CT volumes as structured graphs. In this formulation, each node corresponds to a triplet of axial slices, and edges encode spatial dependencies parameterized by inter-slice spacing along the z-axis. Slice-level features interact through spectral-domain graph convolutions, enabling efficient modeling of both local anatomical context and global volumetric structure. Spatial awareness is further reinforced through an axial positional embedding. We conduct extensive experiments to analyze the impact of graph topology, edge weighting, and feature aggregation strategies, comparing CT-SSG with both standard neural encoders and domain-specific CT architectures. Comprehensive ablations and transfer studies demonstrate the generality of our formulation, including applications to automated radiology report generation and cross-organ adaptation to abdominal CT scans for multi-label abnormality classification.

To summarize our contributions and key advantages of our work: (1) **CT-SSG**: We propose CT-SSG, a new visual encoder that models a 3D CT volume as a graph of triplet axial slices. To capture spatial dependencies, we introduce a Triplet Axial Slice Positional Embedding, along with an edge-weighting strategy for relative position awareness within a spectral-domain GNN module; (2) **Cross-**

dataset generalization: CT-SSG demonstrates strong cross-dataset generalization, maintaining consistent performance when trained on a public Turkish dataset and evaluated on independent datasets from the United States and France. Additionally, we demonstrate the transferability of CT-SSG's pretrained weights from chest to abdominal CT scans, highlighting its potential as a versatile backbone for a broad range of 3D medical imaging tasks; (3) **Ablation study**: We conduct thorough ablation studies to analyze the impact of model depth, hyperparameter choices, graph topology, and connectivity patterns across different convolutional operators. Additionally, we evaluate models under patient-specific variations, including z-axis translations and voxel intensity perturbations. (4) **Transfer to Report generation**: Beyond multi-label abnormality classification, we evaluate CT-SSG on automated radiology report generation, demonstrating that the learned representations are transferable and effective for related CT-based downstream tasks. (5) **Transfer to Abdominal CT for Abnormality Classification**: Although our primary focus is chest CT, we evaluate CT-SSG representations via linear probing on abdominal CT in a low-data regime. We find that chest-pretrained backbones yield stronger performance than supervised training from scratch when fewer than 3,750 samples are available, highlighting the transferability of our approach across anatomical domains.

2. Related Works

2.1 3D Visual Encoder

3D Convolutional Neural Network. Early advances in both 2D and 3D imaging, spanning natural images and medical modalities, have been predominantly driven by Convolutional Neural Networks (CNNs), which demonstrated strong capabilities in extracting fine-grained visual features (LeCun et al., 2015). CNN-based architectures have been successfully applied to a wide range of tasks, including segmentation (Ilesanmi et al., 2024), classification (He et al., 2015), and image captioning (Kougia et al., 2019), across diverse domains such as medical imaging (Anaya-Isaza et al., 2021), earth observation (Bianchi and Barfoot, 2021), and sports analytics (Chang et al., 2024).

3D Transformer Neural Network. Despite their success, CNNs inherently struggle to capture long-range dependencies due to their limited receptive fields, which can hinder their ability to model contextual information, an essential requirement in 3D imaging for understanding large-scale anatomical structures (Ma et al., 2024). Inspired by breakthroughs in Natural Language Processing (Devlin et al., 2018), Vision Transformers (ViTs) were introduced for 2D visual modalities, offering an alternative that leverages self-attention mechanisms (Vaswani et al., 2023) to model global context by enabling interactions between distant image regions (Dosovitskiy et al., 2021). Building upon these principles, Vision Transformers have been extended to 3D data, including applications in video analysis and 3D medical imaging (Wang, 2023). Notably, ViViT, an adaptation of the Vision Transformer for video sequences, applies a Spatial Transformer to model interactions among spatial tokens for each temporal step, followed by a Temporal Transformer to capture dependencies along the temporal axis (Arnab et al., 2021). In the context of CT imaging, ViViT has further inspired frameworks for synthetic volume generation (Hamamci et al., 2023) and automated clinical report synthesis (Hamamci et al., 2024). Similarly, the Swin Transformer, initially designed for 2D vision tasks (Liu et al., 2021b), introduces a hierarchical architecture with shifted windows that enables local and global context modeling while efficiently handling large variations in the scale of visual entities. Swin Transformer was adapted to 3D modalities for various tasks such as video understanding (Liu et al., 2021c), indoor scene understanding (Yang et al., 2023) and organs segmentation of 3D medical images (Tang et al., 2022).

2.5D Neural Network. While Vision Transformers excel at modeling long-range dependencies, they often require extensive pre-training on large-scale, domain-specific datasets to achieve competitive performance, a limitation in medical imaging where annotated datasets are comparatively scarce (Hamamci et al., 2023). A widely adopted alternative is transfer learning from models pre-trained on large-scale natural image datasets (Zhang et al., 2023). In 3D Chest CT imaging, CT-Net was among the first approaches to propose a 2.5D strategy, representing volumetric CT data as stacks of axial slices (Draelos et al., 2021). Feature maps are extracted from each slice using a 2D image encoder pre-trained on natural images, then aggregated through a lightweight 3D convolutional network to produce a compact volumetric representation. This idea was further extended by CT-Scroll, which introduced a hybrid scheme wherein the volume is represented as a set of visual tokens, each associated with triplets of slices (Di Piazza et al., 2025a). These tokens interact through attention mechanisms and are subsequently aggregated via mean pooling. While 3D approaches, such as ViTs or Swin Transformers, incorporate spatial awareness through positional embeddings (Dosovitskiy et al., 2021) or relative positional biases (Liu et al., 2021b), 2.5D methods lack explicit or implicit modeling of spatial continuity within the volume. This limitation may constrain their ability to effectively capture both short- and long-range spatial dependencies.

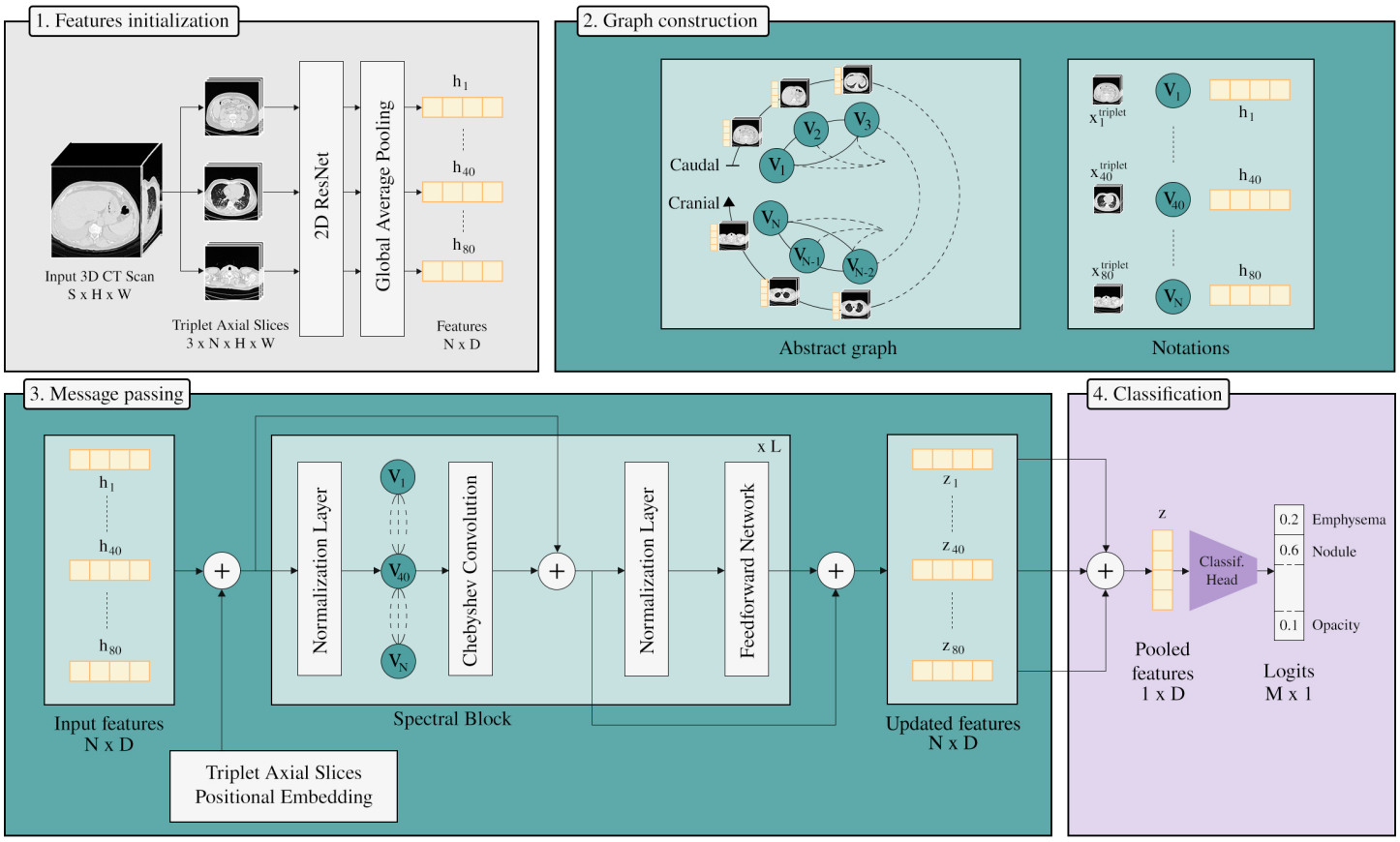

Figure 2: CT-SSG Architecture Overview. Adjacent axial slices are grouped into triplets, each representing a node in a graph. Edges between nodes are weighted according to their physical distance along the z-axis. Node features are enhanced with Triplet Axial Slices positional embeddings, and then processed by a Spectral Block that incorporates Chebyshev graph convolution for structured spectral modeling. The resulting node representations are aggregated via mean pooling and passed to a classification head to predict abnormalities.

2.2 Foundation Models for CT Imaging

Recently, large-scale foundation models have emerged as powerful feature extractors for medical imaging, including CT volumes. Approaches such as CT-CLIP (Hamamci et al., 2026), CT-FM (Pai et al., 2025), Merlin (Blankemeier et al., 2024) and COLIPRI (Wald et al., 2026) leverage large-scale pretraining to learn transferable representations that perform well when transfered to more specific task under limited fine-tuning. Such frameworks often rely on self-supervised objectives including visual reconstruction (Hamamci et al., 2023), visual contrastive learning (Pai et al., 2025) or vision-language alignment (Wald et al., 2026; Blankemeier et al., 2024; Hamamci et al., 2026) from large-scale datasets. Similarly, general-purpose self-supervised models such as DINOv2 (Oquab et al., 2024), that trains a 1B parameters Vision Transformer model on large-scale curated data from diverse sources, have been adapted to medical images to obtain strong slice-level embeddings, both for CT and MRI (Müller-Franzes et al., 2025), as well as for multi-modal patch embeddings from multi-sequence MRI (Scholz et al., 2025). While these methods excel at learning generic visual representations, the visual encoders typically operate within Euclidean framework and do not explicitly model structured inter-slice relationships. In contrast, our work focuses on structured representation learning for CT volumes, a paradigm which can be complementary to foundation models. Notably, CT-SSG can incorporate features extracted from any 2D pretrained encoder, while providing a structured graph reasoning layer tailored to volumetric medical data.

2.3 Graph Neural Network

In various application domains such as biology (Reiser et al., 2022) or transportation (Makarov et al., 2024), graphs are a common representation of data found in nature (Veličković, 2023). A graph, denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consists of a set of edges $\mathcal{E}$ which model the connections between a set of nodes $\mathcal{V}$. In deep learning, GNNs have become the main approach for tasks involving graph-structured data (Bechler-Speicher et al., 2024), where each node is associated with a vector representation, which is iteratively updated through neighborhood aggregation during the forward message passing

process.

Representative models mainly include Convolutional GNNs (GraphConv), which aggregate neighboring node features through graph-based convolutions (Defferrard et al., 2017) or Attentional GNNs (GAT), which leverage attention mechanisms to weight the importance of different neighbors during aggregation (Veličković et al., 2018). Inspired by the attention mechanism (Bahdanau et al., 2016) and self-attention mechanism of the Transformer (Vaswani et al., 2023), the motivation of Graph Attention is to compute a representation of every node as a weighted average of its neighbors (Brody et al., 2022). While spatial networks, including Graph Convolution and Graph Attention, define graph convolution as a localized averaging operation with learned weights, spectral networks define convolution via eigen-decomposition of the graph Laplacian (Zhang et al., 2021). In such spectral networks, the convolution operator is defined in the Fourier domain through localized spectral filters.

In medical imaging, GNNs have been used in tasks such as medical knowledge integration in 2D X-ray radiology report generation (Liu et al., 2021a) to incorporate prior knowledge as a graph of connected textual medical concepts, and Whole Slide Image (WSI) analysis (Guo et al., 2023) to model the hierarchical structure of the pyramids WSI. In the context of Computed Tomography, recent work (Kalisch et al., 2025) models CT volumes as graphs by grouping patches based on anatomical segmentation for report generation. In contrast, our method is purely data-driven, requiring no segmentation labels and is therefore applicable in settings without anatomical annotations. For clarity, this study is restricted to segmentation-free paradigms, differentiating it from anatomical segmentation-based graph methods. Separately, multi-view graph representations have been explored in 3D medical imaging, where each node encodes features from orthogonal axial, sagittal, and coronal slices using a frozen 2D Vision Transformer (Kiechle et al., 2024).

Building on the empirical success of 2.5D approaches, we propose a principled formulation of 3D CT volumes as structured graphs of axial slices. This perspective unifies slice-level representations with inter-slice dependencies, enabling systematic investigation of graph topologies, edge-weighting schemes, and aggregation mechanisms. This work formalizes CT interpretation within a graph-based framework, providing a flexible and general paradigm that bridges 2D and volumetric modeling.

3. Method

As shown in Figure 2, CT-SSG models the 3D CT Scan as a graph of *triplet axial CT slices* with undirected edges weighted by their *physical distance along the caudal-cranial axis*. Node features interact through a spectral domain module, before being pooled and given to a classification head to predict abnormalities. A comprehensive PyTorch pseudocode table outlining each operation, its semantic role, and corresponding tensor shapes is provided in Appendix 13.

Symbol	Description	Value
S	Number of axial slices	240
H_s	Height of an axial slice	480
W_s	Width of an axial slice	480
N	Number of triplet axial slices	80
C	Number of axial slices per triplet	3
d	Dimension of latent space	512
L^*	Spectral module depth	1
q_l^*	Receptive field for layer l	16
K^*	Chebyshev filter size	3

Table 1: Summary of key notations and optimal experimental settings. Symbols marked with * denote tuned hyperparameters.

3.1 Notations

We consider a multi-label abnormality classification task with an input space $\mathcal{X} \in \mathbb{R}^{S\times H_s\times W_s}$ and a target space $\mathcal{Y} \in [1, \cdots, M]$. S refers to the number of axial slices, each of dimension $H_s \times W_s$. M is the number of abnormalities. Table 1 details each variable with description and corresponding value for experiments.

3.2 Features Initialization

Following a 2.5D strategy, we partition the input volume X into N non-overlapping triplets of slices to encode minimal local 3D context while maintaining compatibility with 2D backbones, resulting in a tensor of shape $N \times C \times H_s \times W_s$. Using three adjacent slices allows the network to capture inter-slice anatomical continuity and subtle volumetric context without resorting to full 3D convolutions or excessively large context windows that may dilute localized patterns. Each triplet, noted as $x_i^{\text{triplet}} (i \in \{1, \ldots, N\})$, is processed independently by a learnable 2D ResNet-18 backbone (He et al., 2015), extracting spatial features. The N output features maps are subsequently aggregated via mean pooling to produce a compact representation for each slice triplet. This features initialization step maps each triplet of axial slice into a d-dimensional embedding, denoted as $\bar{h}_i \in \mathbb{R}^d$. The use of ResNet-18 provides a favorable trade-off between representational capacity and computational efficiency while allowing us to focus on evaluating the contribution of the proposed graph-based feature aggregation. Importantly, the proposed CT-SSG formulation is backbone-agnostic.

To empirically validate this property, we additionally evaluate a higher-capacity ResNet-34 encoder in Appendix H, demonstrating consistent performance scaling without modification to the graph architecture.

3.3 Triplet Axial Slices Positional Embeddings

After obtaining all triplet slices embeddings $\bar{H} = [\bar{h}_1, \ldots, \bar{h}_N]$, triplet axial slices positional embeddings are added to retain positional information along the caudal-cranial axis. We use learnable 1D position embeddings, denoted as $P_{\text{pos}}^{\text{axial}} \in \mathbb{R}^{N \times d}$, resulting as a sequence of embedding vectors H, such that:

$$H = \bar{H} + P_{\text{pos}}^{\text{axial}} \,. \tag{1}$$

3.4 Graph Construction

We define the volumetric representation as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, H, A)$. In this section, we define nodes, edges, node features and the adjacency matrix.

Nodes. $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is the set of nodes, where each node v_i represents a triplet of 3 consecutive axial slices.

Edges. $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, where an edge $(v_i, v_j) \in \mathcal{E}$ is weighted based on a function of inter-triplet distance and z-axis spacing. The weight of an edge (v_i, v_j), denoted as $w_{i,j} \in \mathbb{R}^+$, is defined such that:

$$w_{i,j} = 1 + \frac{1}{1 + 3 \times |i - j| \times s_z} \,, \tag{2}$$

where s_z is the spacing along the caudal-cranial axis in millimeters.

We further investigate the impact of graph connectivity by exploring a family of topologies parameterized by a receptive field size $q \in \mathbb{N}^+$. Specifically, we construct an undirected edge $(v_i, v_j) \in \mathcal{E}$ between nodes if their corresponding triplet slices are at most q steps apart in the sequence, yielding the edge set:

$$\mathcal{E} = \{(v_i, v_j) \mid |i - j| \leq q\} \,. \tag{3}$$

In Section 5.2, we perform a comprehensive ablation study to assess how varying q influences the performance of different GNN architectures, highlighting the role of graph receptive field in modeling caudal-cranial axis dependencies within 3D CT volumes.

Nodes features. $H = \{h_1, \ldots, h_N\} \in \mathbb{R}^{N \times d}$ is the node feature matrix, where $h_i \in \mathbb{R}^d$ denotes the feature embedding of node v_i.

Adjacency matrix. $A \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix, where $A_{ij} = w_{i,j} \in \mathbb{R}^+$ encodes the connectivity and spatial relationship between triplets, $w_{i,j}$ being the edge weight such that:

$$A_{ij} = \begin{cases} w_{ij}, & \text{if } (v_i, v_j) \in \mathcal{E} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

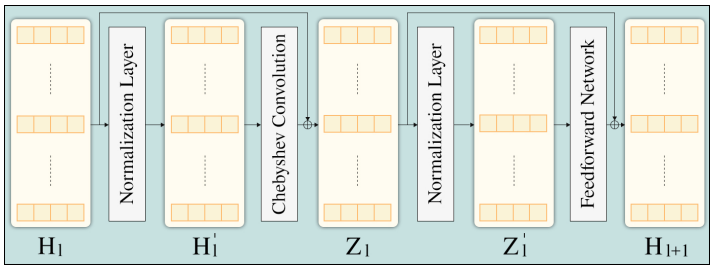

Figure 3: Spectral Block with detailed notations. Input features are given to a first normalization layer, followed by spectral graph convolutions with a residual skip connection. These updated features are then fed to a feedforward neural network followed by a second normalization layer with a residual skip connection.

3.5 Spectral Domain Module

A key challenge in this formulation is the variability in anatomical positioning across patients due to differences in scan length and body proportions. Traditional spatial graph convolutions, such as GraphConv (Morris et al., 2021), aggregate information from fixed local neighborhoods, which can be suboptimal in this context as anatomical structures do not consistently align across scans. Instead, we leverage Chebyshev convolutions (Defferrard et al., 2017) to define graph convolutions in the spectral domain, each followed by a feedforward neural network. Unlike spatial approaches, which struggle with non-uniform neighborhood structures (Bruna et al., 2014), ChebConv utilizes polynomial approximations of the graph Laplacian (Belkin and Niyogi, 2001) to capture hierarchical feature representations while preserving spatial localization. This allows the model to adapt to variations in caudal-cranial slice positioning and effectively learn long-range anatomical relationships, making it more robust to inter-patient variability.

We introduce a Spectral Module, denoted as Φ_{SM}, consisting L Spectral Blocks. Each block consists of two sublayers. While Figure 2 shows the overall CT-SSG architecture, Figure 3 presents a detailed schematic of the spectral block, where all operations and symbols are explicitly annotated to facilitate interpretation of the notation.

The first sublayer consists of a Normalization Layer (Ba et al., 2016), noted f_l^{LN}, and followed by a spectral convolution. Specifically, we leverage a Chebyshev Convolution, denoted as f_l^{Cheb}, to benefit from its polynomial formulation that allows us to capture information from K-hop

neighborhood. Let $H_0 = H$, and H_l denote the input features of the l-th block, we formaly define the forward pass in the first sublayer such that:

$$Z_l = H_l + \left(f_l^{\text{Cheb}} \circ f_l^{\text{LN}}\right)(H_l)\,. \tag{5}$$

For the Chebyshev convolution, denoted as f_l^{Cheb}, the scaled and normalized Laplacian $\hat{L}$ is defined as:

$$\hat{L} = \frac{2}{\lambda_{\text{max}}}(D - A) - I\,, \tag{6}$$

where λ_{max} is the largest eigenvalue of the graph Laplacian $L = D - A$. The degree matrix D is a diagonal matrix where $D_{i,i} = \sum_{j=1}^{N} w_{i,j}$. The convolution operation is parameterized using Chebyshev polynomials $T_j(\hat{L}) \in \mathbb{R}^{N\times N}$, resulting in a recurrence relation for the transformation of the node feature matrix. Let $\theta_{l,k} \in \mathbb{R}^{d\times d}$ be the learnable parameters, and K be the Chebyshev filter size. The recurrence relation is given by:

$$f_l^{\text{Cheb}}(X) = \sum_{k=0}^{K-1} T_{l,k}(\hat{L}) X \theta_{l,k}\,. \tag{7}$$

We investigate the effect of the filter size K on model performance in the ablation study presented in Section 5.2. The Chebyshev convolution was implemented with ChebConv module from PyTorch Geometric. Spectral graph convolutions are defined with respect to the graph Laplacian and are therefore tied to the underlying graph topology. In our setting, this dependency is mitigated by the use of a fixed and regular graph structure reflecting the axial organization of CT volumes, which is shared across all samples. We note that extending the proposed approach to substantially different or data-dependent graph topologies may require adapted spectral formulations.

The second sublayer consists of another Normalization Layer, denoted as g_l^{LN}, followed by a feedforward neural network, noted g_l^{FNN} and implemented as a linear layer followed by a GELU activation function (Shazeer, 2020). The second sublayer is also followed by a residual connection, as followed:

$$H^{l+1} = Z^l + \left(g_l^{\text{FFN}} \circ g_l^{\text{LN}}\right)(Z^l)\,. \tag{8}$$

Formally, the Spectral Module outputs updated features, denoted as $Z = H_L = [z_1, \ldots, z_N]$ with $z_i \in \mathbb{R}^d$ being the updated features for the i-th node, such that:

$$Z = \Phi_{\text{SM}}(H)\,. \tag{9}$$

3.6 Classification

The obtained vector representations are aggregated through mean pooling to derive a vector representation, denoted as $\bar{z} \in \mathbb{R}^d$, such that:

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i\,. \tag{10}$$

$\bar{z}$ is subsequently passed to a classification head, noted Ψ, which predicts the logit vector $\hat{y} \in \mathbb{R}^M$. The model is trained on a multi-label classification task using Binary Cross-Entropy as the loss function (Mao et al., 2023).

4. Dataset

Databases. All models are trained using 5-fold cross-validation (Stone, 1974) and evaluated on the `CT-RATE` dataset, which consists of non-contrast chest CT volumes annotated with 18 abnormalities from 21,304 unique patients (Hamamci et al., 2026). These labels are automatically extracted from radiology reports using RadBERT (Yan et al., 2022), a language model trained to extract abnormalities for radiology report. Duplicates are not retrained during training. To assess cross-dataset generalization, models are also evaluated on the external `RAD-ChestCT` test dataset, using the 16 abnormalities shared with `CT-RATE` from 1,334 unique patients, which are extracted from reports via a SARLE-based labeler (Draelos et al., 2021). Additionally, the `CT-HCL` internal dataset comprises non-contrast chest CT scans from 2,000 unique adult patients from the Hospices Civils de Lyon, with 9 abnormalities shared with `CT-RATE`. These labels are manually extracted from radiology reports by radiologists (Jupin-Delevaux et al., 2023). For cross-dataset evaluation databases (`RAD-ChestCT` and `CT-HCL`), exact abnormality labels do not perfectly align with those in `CT-RATE`. To address this, we map related abnormalities into broader semantic groups (e.g., both *Artery wall calcification* and *Coronary artery wall calcification* are grouped under *Calcification*). At inference time, following the protocol of the `CT-RATE` original paper, the model's prediction for each abnormality group is derived by taking the maximum predicted probability among all constituent abnormalities within that group (Hamamci et al., 2026). This approach enables a consistent comparison across datasets despite label granularity differences. Figure 4 provides a comprehensive comparison of the test sets from `CT-RATE`, `RAD-ChestCT` and `CT-HCL` datasets.

Processing. Consistent with prior work (Hamamci et al., 2024; Draelos et al., 2021; Di Piazza et al., 2025a), all datasets are processed following the same pipeline to ensure fair evaluation. CT scans are reformated to a SLP orientation, such that first axis points from inferior to superior, the second from right to left, and the third from interior to superior. Volumes are cropped or padded to a standardized resolution of $240 \times 480 \times 480$, with a spacing of 0.75 mm along the z-axis and 1.5 mm along the x- and y-axes (z,

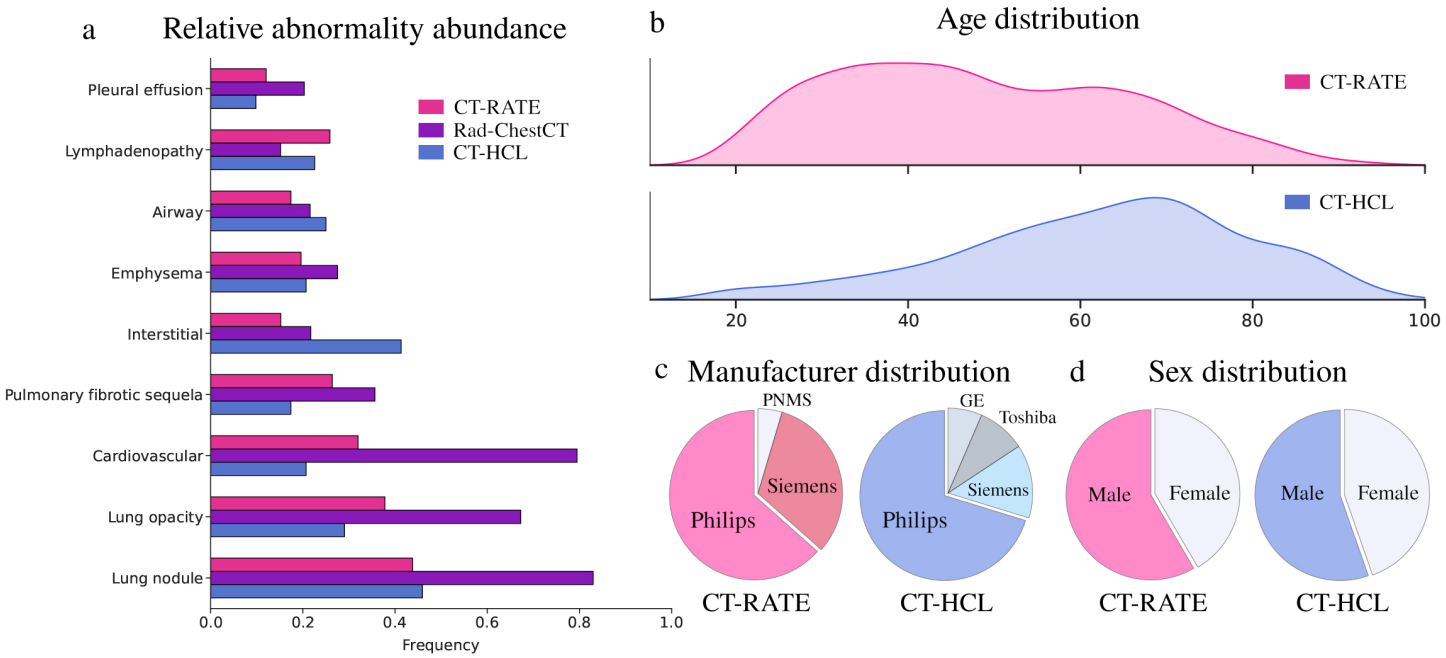

Figure 4: Comprehensive analysis of the datasets. Metadata not available for the `RAD-ChestCT` dataset. a) Abnormalities from `CT-HCL` are extracted with a BERT-based language model trained on french radiology reports from manually extracted anotations. b) `CT-HCL` comprises data from 2,000 unique patients, with age randing from 20 to 100 years. c) `CT-HCL` volumes comes from Hospices Civil de Lyon, with scanners from four manufacturers. d) `CT-HCL` volumes were acquired both from male and female patients.

x, y). Hounsfield Units are clipped to the range [-1000, 200], which corresponds to the practical diagnostic window (DenOtter and Schubert, 2024), before being mapped to the range $[0, 1]$ and normalized using ImageNet statistics (-0.449) (Russakovsky et al., 2015).

5. Experiments

Our experimental results comprise 4 sections: (Section 5.1) We provide quantitative and qualitative results on the multi-label abnormality classification task; (Section 5.2) We perform an ablation study on CT-SSG components; (Section 5.3) We evaluate model's robustness to patient body translations along the z-axis and to intensity noise perturbations; (Section 5.4) We extend results to the report generation task; and (Section 5.5) we evaluate our model's transfer learning ability on CT abdominal scans for abnormality classification.

5.1 Multi-label abnormality Classification

Baselines. We compare our approach against three categories of baselines. First, we consider 3D Convolutional Neural Networks (■) with an inflated ResNet3D (Carreira and Zisserman, 2018), which which leverages filter inflation from pretrained 2D networks to learn transferable volumetric representations. Second, we evaluate against Transformer-based architectures (■), including ViT3D, a straightforward extension of Vision Transformer to 3D volumes (Dosovitskiy et al., 2021); ViViT (Arnab et al., 2021), originally designed for video processing; and Swin3D, an adaptation of the Swin Transformer with hierarchical window-based attention for 3D inputs (Liu et al., 2021c). Third, we benchmark against 2.5D methods (■), which process 2D slices to extract feature maps. This includes CT-Net, which aggregates triplet axial slices features using a lightweight 3D CNN (Draelos et al., 2021), and CT-Scroll, which employs alternating local and global attention mechanisms to capture dependencies across slices (Di Piazza et al., 2025a). Additionally, we include a Multi-View Graph (CT-MvG) baseline, which represents each 3D volume as a graph of nodes corresponding to orthogonal axial, sagittal, and coronal slices (Kiechle et al., 2024).

To ensure a fair comparison, all models are initialized with ImageNet-pretrained weights, either directly for 2D architectures or via weight inflation (Zhang et al., 2023) for 3D counterparts, promoting stable and efficient convergence. Specifically, the ResNet3D via weight inflation (Carreira and Zisserman, 2018) from a 2D ResNet-18 (He et al., 2015) pretrained on ImageNet (Russakovsky et al., 2015); ViT3D and ViViT via weight inflation from a 2D ViT-S16 (Dosovitskiy et al., 2021) pretrained on ImageNet; and Swin3D via weight inflation from a 2D Swin-S16 (Liu et al., 2021b) pretrained on ImageNet. Since Vision Transformers families of models are typically release in multiple capacity vari-

Table 2: Performance of the models trained and evaluated on the CT-RATE dataset. Mean and standard deviation are computed across 5 cross-validation folds. Experiments with (†) refer to a cross-dataset evaluation from models trained on CT-RATE, and assessed on the RAD-ChestCT and CT-HCL datasets. Random refer to predictions sampled from a uniform distribution. **Best** results are in bold, second-best are underlined. 3D Transformer, 3D CNN, 2.5D.

Dataset	Method	F1-Score	AUROC	Accuracy	mAP
CT-RATE	Random	27.78±0.51	49.88±0.62	49.89±0.31	20.94±0.12
	ViT3D (Dosovitskiy et al., 2021)	49.53±0.51	78.97±0.37	75.37±0.52	46.31±0.43
	ViViT (Arnab et al., 2021)	50.66±0.45	80.03±0.18	76.81±0.81	48.57±0.66
	Swin3D (Liu et al., 2021c)	50.25±0.24	79.39±0.37	76.63±0.89	47.26±0.35
	ResNet3D (Carreira and Zisserman, 2018)	51.51±0.48	77.48±0.64	80.67±0.60	50.81±0.71
	CT-Net (Draelos et al., 2021)	51.04±0.60	79.69±0.27	77.78±0.19	48.52±0.64
	CT-MvG (Kiechle et al., 2024)	52.35±0.17	81.91±0.16	78.25±0.46	51.51±0.46
	CT-Scroll (Di Piazza et al., 2025a)	54.30±0.20	82.18±0.27	79.68±0.67	53.09±0.48
	CT-SSG (Ours)	**57.18**±0.19	**83.64**±0.21	**81.03**±0.38	**58.24**±0.50
RAD-ChestCT(†)	Random	35.91±0.41	49.68±0.55	50.40±0.32	32.65±0.08
	ViT3D (Dosovitskiy et al., 2021)	47.94±0.85	66.85±0.18	61.42±1.90	47.94±0.32
	ViViT (Arnab et al., 2021)	49.23±0.44	68.89±0.24	62.08±0.76	49.40±0.40
	Swin3D (Liu et al., 2021c)	47.36±1.10	66.34±0.45	61.34±1.41	47.71±0.66
	ResNet3D (Carreira and Zisserman, 2018)	48.56±1.29	69.94±0.92	61.04±0.94	51.37±0.74
	CT-Net (Draelos et al., 2021)	49.09±0.83	69.09±0.48	62.32±0.64	49.66±0.37
	CT-MvG (Kiechle et al., 2024)	50.33±1.05	71.19±0.43	62.84±0.30	52.82±0.69
	CT-Scroll (Di Piazza et al., 2025a)	49.47±0.82	71.43±0.27	63.86±1.87	52.53±0.52
	CT-SSG (Ours)	**52.25**±0.88	**74.58**±0.36	**69.37**±1.74	**58.75**±0.28
CT-HCL(†)	Random	33.38±0.28	50.16±0.41	49.98±0.40	27.01±0.13
	ViT3D (Dosovitskiy et al., 2021)	44.77±0.39	64.77±0.40	52.17±0.93	41.94±0.16
	ViViT (Arnab et al., 2021)	45.98±0.97	67.08±0.52	55.16±1.30	43.48±2.19
	Swin3D (Liu et al., 2021c)	44.76±0.99	64.41±0.37	52.94±0.74	41.73±0.42
	ResNet3D (Carreira and Zisserman, 2018)	45.92±0.42	67.95±1.02	53.17±0.65	45.81±0.70
	CT-Net (Draelos et al., 2021)	46.81±0.83	66.53±0.49	54.70±0.33	43.16±0.43
	CT-MvG (Kiechle et al., 2024)	47.10±0.52	69.04±0.33	56.13±1.38	47.10±0.52
	CT-Scroll (Di Piazza et al., 2025a)	48.24±0.53	69.36±0.27	56.88±1.81	46.46±0.39
	CT-SSG (Ours)	**50.26**±0.57	**71.77**±0.36	**61.43**±1.18	**51.81**±0.44

ants, we adopt the *Small* variants across all baselines to ensure comparability. This consistent choice offer comparable parameter counts and computational budgets, enabling a balanced comparison without favoring a particular design and representing a practically deployable setting while still retaining sufficient capacity to serve as strong baselines. Similarly, 2.5D models leveraging ResNet-18 backbones use ImageNet-pretrained 2D ResNet-18 weights at initialization. The 2D ViT-S16 module within CT-MvG is initialized with ImageNet-pretrained weights. Both for CT-SSG and baselines, all parameters are trainable during training to ensure a fair and consistent comparison across methods. Appendix 15 reports the training details of all methods.

Evaluation protocol. All models are trained and evaluated using 5-fold cross-validation. For each run, we apply early stopping based on the checkpoint that achieves the highest macro F1-score on the validation set, i.e. the harmonic mean of precision and recall, averaged across all abnormalities. We report performance metrics on the test set corresponding to this selected checkpoint.

Quantitative Results. Table 2 reports the quantitative performance of all methods in terms of macro F1-Score, AUROC, Accuracy, and mean Average Precision (mAP). On the CT-RATE test set, CT-SSG achieves a macro-averaged F1-Score of 57.18, yielding relative gains of $+\Delta 5.30\%$ over CT-Scroll (Di Piazza et al., 2025a), $+\Delta 11.01\%$ over ResNet3D (Anaya-Isaza et al., 2021) and $+\Delta 12.87\%$ over ViViT (Arnab et al., 2021). Paired t-tests across all metrics indicate $p - value < 0.01$, confirming that the improvements are statistically significant for $\alpha = 0.01$, α being

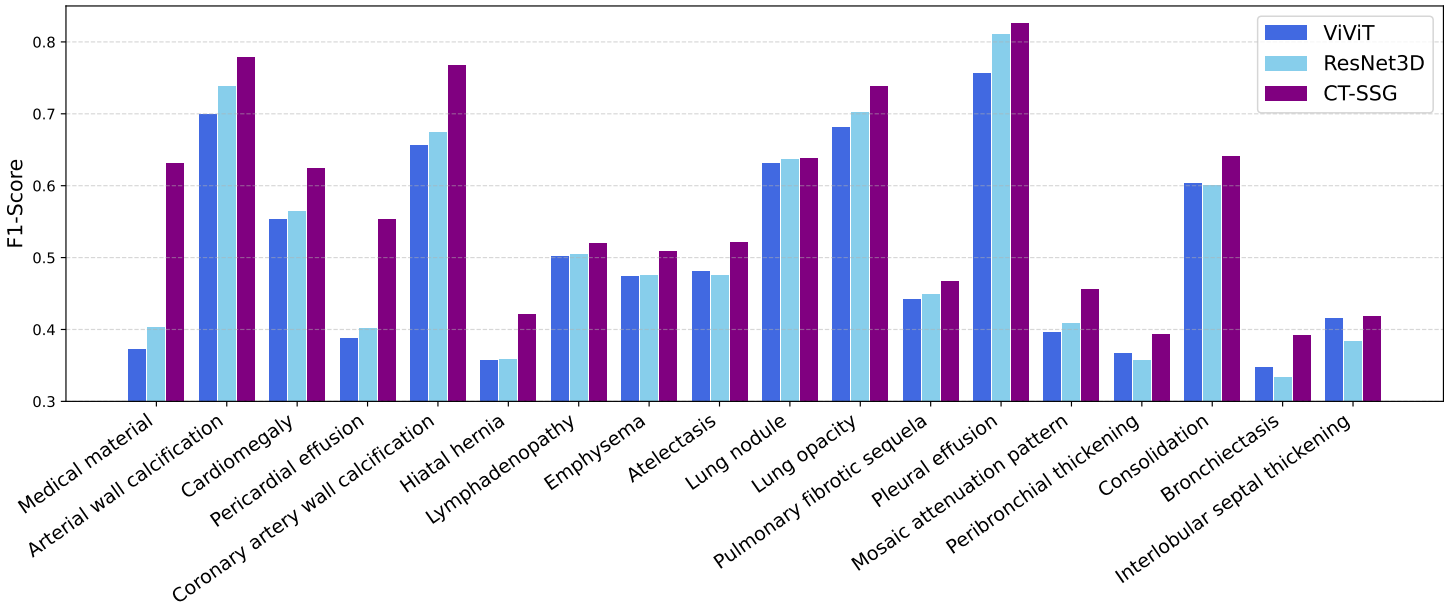

Figure 5: F1-Score per abnormality for the 18 abnormalities from the `CT-RATE` test set, comparing our proposed CT-SSG with representative 3D Convolutional and 3D Transformer baselines. For clarity, one representative model per family is reported. CT-SSG consistently improves over both baselines, with the largest absolute gains observed in *Pericardial effusion* ($+\Delta 8.96\%$), *Calcification* ($+\Delta 6.23\%$), and *Pleural effusion* ($+\Delta 6.20\%$).

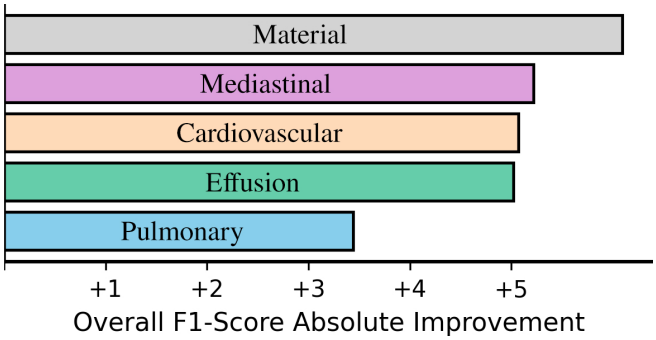

Figure 6: Average absolute F1-score improvements of CT-SSG over representative 3D convolutional (CNN) and 3D Transformer (ViViT) baselines. Abnormalities are grouped by anatomical region and pathophysiological type to highlight systematic patterns of gain. CT-SSG yields consistent improvements across groups.

the Type I error rate (Ross and Willson, 2017). In cross-dataset evaluations on `RAD-ChestCT` and `CT-HCL`, CT-SSG consistently ranks highest across metrics, indicating strong generalization to distinct clinical distributions. Although the harmonization of the abnormality taxonomy may affect absolute scores, the relative ordering of methods is preserved across metrics and datasets.

Figure 6 reports the average absolute F1-score gain of CT-SSG over ResNet3D and ViViT, aggregated by anatomical region and pathophysiological category. Notably, CT-SSG yields a $+\Delta 3.2\%$ improvement for pulmonary diseases and $+\Delta 5.0\%$ for mediastinal, cardiovascular, and effusion diseases, indicating that the performance gains are consistent across diverse abnormality types. Figure 5 presents per-abnormality F1-scores for CT-SSG, ViViT, and ResNet3D, while Table 10 details the results for all baselines, showing that CT-SSG achieves superior classification performance for the majority of abnormalities. Abnormality-wise performance and t-distributed stochastic neighbor embedding (t-SNE) projections produced by CT-SSG are provided in Appendix A and Appendix D, respectively.

Contextualization with CT Foundation Models. We additionally contextualize our approach against recent CT-specific foundation models. We include CT-CLIP pretrained with vision-language alignment from chest CT (Hamamci et al., 2026), Merlin also pretrained with vision-language alignment from abdominal CT (Blankemeier et al., 2024), CT-FM pretrained on CT scans with a wide variety of body parts through visual contrastive learning (Pai et al., 2025), and COLIPRI pretrained on chest CT with both vision-only, vision-language alignment, opposite sentences and report generation objectives (Wald et al., 2026). As these foundation models are pretrained using objectives different from supervised abnormality classification, we evaluate them under a linear probing protocol. Specifically, the visual encoder of each pretrained model is frozen, and only a classification head is trained for multi-label abnormality prediction. Training splits and evaluation protocols are kept consistent across all methods to ensure a controlled comparison. We note, however, that the pretraining data of CT-CLIP and COLIPRI partially overlap with our validation

sets used for linear probing, as these models are pretrained on the full CT-RATE train set. Importantly, all reported results are evaluated on strictly held-out test sets for both CT-RATE and RAD-ChestCT, with no overlap with either pretraining or linear probing data.

Method		CT-RATE		RAD-ChestCT	
Setup	Model	AUROC	mAP	AUROC	mAP
Foundation Probing	CT-FM	74.35±0.09	39.22±0.19	63.26±0.13	43.78±0.15
	CT-CLIP(*)	76.85±0.04	42.87±0.08	63.38±0.05	44.30±0.06
	Merlin	77.85±0.04	46.53±0.07	69.80±0.05	49.84±0.08
	COLIPRI(*)	**84.01**±0.03	57.38±0.08	73.98±0.02	55.45±0.10
Supervised	CT-SSG	83.64±0.21	**58.24**±0.50	**74.58**±0.36	**58.75**±0.28

Table 3: Comparison of linear probing on frozen visual encoders from foundation models with CT-SSG, on CT-RATE. **Best results** in bold, second-best underlined. (*)Foundation models are pretrained on data that partially overlap with ours CT-RATE validation splits. However, all reported evaluations are conducted on strictly held-out test sets, with no overlap for either CT-RATE or RAD-ChestCT.

Table 3 shows that CT-SSG outperforms CT-FM, CT-CLIP and Merlin on CT-RATE, in terms of both AUROC and mAP, achieving the best mAP among all compared methods. While COLIPRI attains the highest AUROC on CT-RATE, CT-SSG achieves a closely comparable performance. On RAD-ChestCT, CT-SSG yields the best results in both AUROC and mAP when compared to CT-specific foundation models (paired t-test across folds, $p < 0.01$). These results indicate that CT-SSG's learned representations are well suited for multi-label abnormality analysis and generalize well to unseen datasets, even when compared with representations derived from foundation models with substantially higher architectural complexity trained using generic large-scale pretraining objectives. Additional details on the evaluated CT foundation models, including pretraining objectives, data processing, model complexity and feature extraction protocols, are provided in Appendix E.

Qualitative Results. In addition to the quantitative analysis, Figure 7 presents qualitative examples of correct predictions for each abnormality in the CT-RATE dataset. We visualize Gradient-Weight Class Activation Mapping (Selvaraju et al., 2019) heatmaps, where darker regions correspond to lower activations. Specifically for each volume, we obtain a heatmap of shape $S \times H_s \times W_s$ and we display the s-th axial slice with highest activation, highlighting CT-SSG's ability to classify abnormalities from relevant regions.

5.2 Ablation Studies

Impact of each model component. Table 4 quantifies the incremental impact of each component. Starting from the initialization of the node features, we progressively integrate seven architectural elements. Each addition results in a positive improvement in both F1-score and AUROC. For example, the integration of spectral network yields a $+\Delta 2.18\%$ improvement in F1-Score over spatial convolution, while the axial positional encoding module results in a $+\Delta 0.39\%$ improvement. The cumulative trend underscores that the model benefits from the synergistic effect of multiple design choices, which suggests that each component addresses complementary aspects of the task and collectively ensures robustness.

Component	AUROC	F1-Score	ΔF1 (abs)	ΔF1 (%)
Triplet slices embeddings	81.20±0.42	52.41±0.69	N/A	N/A
+ Spatial convolution	82.45±0.19	54.25±0.37	+1.84	+3.51
+ Spectral convolution	83.06±0.09	55.43±0.44	+1.18	+2.18
+ Normalization layer	83.28±0.19	56.16±0.39	+0.73	+1.32
+ Residual connection	83.32±0.16	56.37±0.13	+0.21	+0.37
+ Edge weighting	83.37±0.33	56.54±0.15	+0.17	+0.30
+ Axial. pos. encoding	83.53±0.23	56.76±0.31	+0.22	+0.39
+ Sparse topology	83.64±0.21	57.18±0.19	+0.42	+0.74

Table 4: Incremental contribution of each model component, on the CT-RATE test set. Starting from the base architecture, components are added cumulatively across rows. For each step, we report F1-score and AUROC, along with absolute and relative improvements (ΔF1) over the configuration in the preceding row. All components yield consistent gains, indicating that the overall performance arises from complementary contributions.

Impact of node design. To investigate the impact of slice selection, Table 5 compare a graph construction with triplets of adjacent slices to one with triplets formed by repeating the same slice three times. Using adjacent axial slices for the three input channels yields consistently better performance than repeating a single slice across channels, suggesting that incorporating limited inter-slice context is beneficial for abnormality classification.

Node design	Slice indices	F1-Score	AUROC	mAP
Adjacent slices	$(s, s+1, s+2)$	57.18±0.19	83.64±0.21	58.24±0.50
Single-slice (x3)	(s, s, s)	56.64±0.26	83.34±0.29	57.31±0.25

Table 5: Comparison between adjacent-slice and single-slice triplets replicated across channels. Adjacent triplets consistently outperform replicated single-slice representations, highlighting the benefit of local inter-slice context. Best results are underlined.

Impact of receptive field. We further analyze the impact of the receptive field, defined as the number of neighboring nodes connected to each node, corresponding to the number of axial slice triplets considered for feature aggregation. Figure 8 reports F1-Score and AUROC for receptive field sizes in $\{1, 16, 32, 64, 80\}$, where 80 corresponds to the maximum number of triplets in a CT scan. Higher

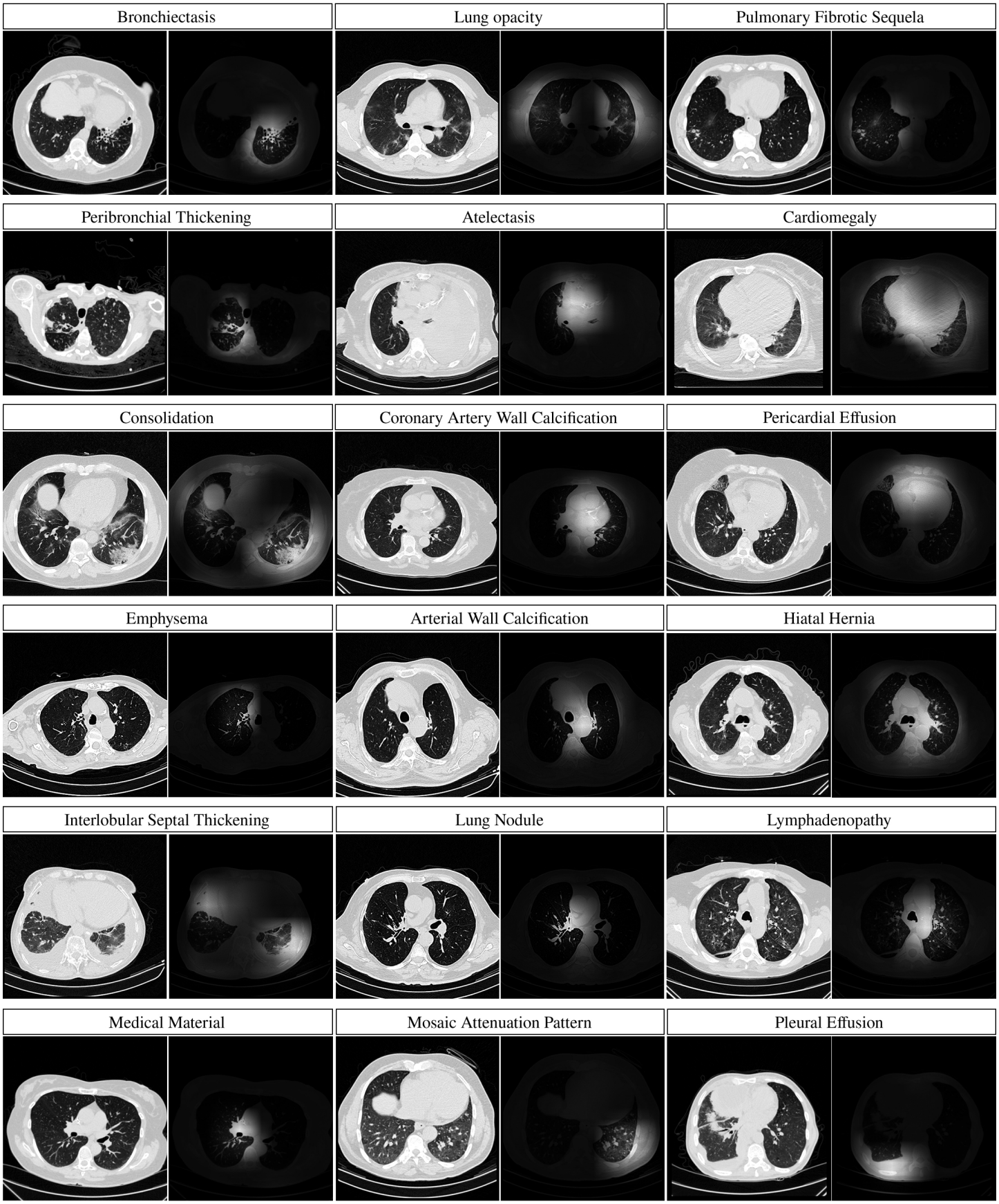

Figure 7: Gradient-weighted class activation maps, extracted from the 2D ResNet from the triplet slices embeddings module, where darker regions indicate lower activations. For each input, we display the slice with the highest absolute activation value from the heatmap.

performance is observed for smaller receptive fields, with peak F1-Score at 16 neighbors in our settings, suggesting that restricting the aggregation to a local neighborhood facilitates effective integration of relevant visual context for abnormality classification.

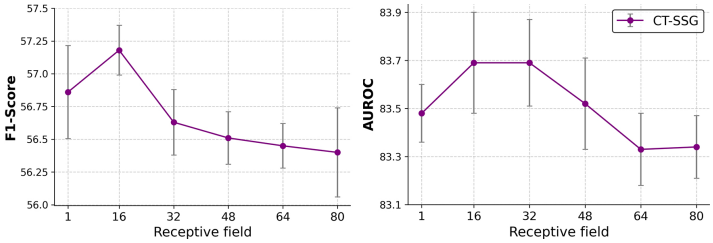

Figure 8: Impact of the receptive field. We compare different receptive field sizes in $\{1, 16, 32, 64, 80\}$, where 80 is the maximum number of nodes. Higher F1-Score and AUROC is observed when restricting the aggregation to a local neighborhood.

Impact of model depth. Table 6 details multi-label abnormality classification performances with 1, 3, and 5 spectral blocks. Increasing the number of propagation layers does not improve performance. In fact, a single layer yields the best results. This indicates that most discriminative information resides in immediate slice-to-slice dependencies, and that effective modeling of 3D CT does not necessarily require deep graph structures, but rather careful design of local inter-slice connectivity.

Model depth L	F1-Score	AUROC	mAP	Accuracy
1	$\underline{57.18}_{\pm 0.19}$	$\underline{83.64}_{\pm 0.21}$	$\underline{58.24}_{\pm 0.50}$	$81.03_{\pm 0.38}$
3	$56.46_{\pm 0.20}$	$83.04_{\pm 0.09}$	$56.58_{\pm 0.35}$	$81.44_{\pm 0.60}$
5	$56.47_{\pm 0.16}$	$83.09_{\pm 0.06}$	$56.79_{\pm 0.25}$	$\underline{82.11}_{\pm 0.51}$

Table 6: Impact of the model depth. We compare models with 1, 3, and 5 layers. Performance peaks at a single layer, suggesting that shallow inter-slice message passing is sufficient for effective representation learning.

Impact of the spectral filter size. Spectral polynomial filters present « two sources of locality ». First, the adjacency matrix defines *who is a neighbor of whom*. For a fully connected graph, every node would be a 1-*hop connected to every other node*. Second, the spectral filter size K which defines *how far the Laplacian power is applied*. For a fully connected graph, every node is already directly 1-hop away from every other node, which means that increasing K does not expand the receptive field but just adds higher-order polynomials of the Laplacian. However if the graph is sparse, the spectral filter size K truly increases the receptive field, making the spectral filter exactly K-localized. Hence, we systematically evaluate the impact of spectral filter size K, which controls the Chebyshev polynomial order in Equation 7, under different graph topology: fully connected and sparse.

Table 7 summarizes the effect of varying the spectral filter size $K \in 1, 3, 5$ across different graph topologies. On the fully connected topology, the best results yields for $K = 5$, but the improvement is not statistically significant at the conventional $\alpha = 0.05$ (paired t-test, $p = 0.077$), though it approaches significance. In contrast, on the sparse graph with receptive field fixed to 16, a larger filter size proves beneficial: $K = 3$ achieves the highest F1-score (57.18), and the improvement over $K = 1$ is statistically significant ($p < 0.01$). These findings highlight the dual role of connectivity and spectral filter size in shaping the effective receptive field of the network. In fully connected graphs, even small Chebyshev orders ($K = 1$) suffice to capture global context, since each node already aggregates information from all others. Increasing K in this setting may lead to redundant propagation and potential oversmoothing, which explains why larger filters do not improve performance. In contrast, under constrained receptive fields (e.g., sparse graphs with fixed neighborhood size), higher-order filters become crucial. A moderate filter size ($K = 3$) expands the receptive field sufficiently to integrate useful multi-hop context while avoiding the noise introduced by excessively large filters. This suggests that spectral filters primarily act as a mechanism to compensate for sparsity in connectivity, while in dense regimes their effect diminishes or even harms performance.

Topology	Filter size K	F1-Score	AUROC	mAP
Fully connected $q = 80$	1	$56.76_{\pm 0.41}$	$83.63_{\pm 0.23}$	$57.56_{\pm 0.47}$
	3	$56.40_{\pm 0.44}$	$83.34_{\pm 0.13}$	$57.39_{\pm 0.31}$
	5	$56.83_{\pm 0.43}$	$83.56_{\pm 0.19}$	$57.69_{\pm 0.32}$
Sparse $q = 16$	1	$56.50_{\pm 0.19}$	$83.63_{\pm 0.21}$	$57.57_{\pm 0.21}$
	3	$\underline{57.18}_{\pm 0.19}$	$\underline{83.64}_{\pm 0.21}$	$\underline{58.24}_{\pm 0.50}$
	5	$56.41_{\pm 0.31}$	$83.37_{\pm 0.19}$	$57.35_{\pm 0.22}$

Table 7: Impact of the spectral filter size K, for different graph topologies. q refers to the receptive field.

Impact of convolutional operator. Table 8 summarizes abnormality classification performance across different graph operators. Among them, the Chebyshev spectral convolution (Defferrard et al., 2017) achieves the strongest results, both on fully connected and sparse topologies. In particular, the spectral model consistently outperforms its spatial counterparts, yielding a relative improvement of $+\Delta 2.82\%$ in F1-Score compared to Graph Attention (Veličković et al., 2018), and $+\Delta 0.65\%$ compared to Graph Convolution (Morris et al., 2021). These results highlight the advantage of leveraging spectral formulations to capture dependencies across slices, and suggest that Graph Convolution may be too restrictive, while Graph Attention that can be more ex-

pressive, often require more parameters and larger training size to be competitive.

Impact of graph topology. Building on this operator-level analysis, we next examine how graph connectivity influences performance. Specifically, we compare a fully connected topology, analogous to the Transformer formulation where all nodes (triplet of axial slices) attend to one another, with a sparse topology that constrains interactions to local neighborhoods along the caudal–cranial axis. Across all operators, Table 8 shows that sparse graphs consistently outperform fully connected ones. On average, sparse topologies yield a $+\Delta 0.82\%$ improvement in F1-Score, suggesting that limiting the receptive field to local interactions better captures short-range dependencies between adjacent slices and ultimately enhances abnormality classification performance.

Network	Operator	Topology	F1-Score	AUROC
Spatial	Graph Conv.	Fully connected	$56.35_{\pm 0.18}$	$83.12_{\pm 0.22}$
		Sparse	$\underline{56.81}_{\pm 0.28}$	$\underline{83.44}_{\pm 0.25}$
	Graph Attention	Fully connected	$55.05_{\pm 0.19}$	$82.81_{\pm 0.07}$
		Sparse	$\underline{55.61}_{\pm 0.16}$	$\underline{82.98}_{\pm 0.18}$
Spectral	Chebyshev	Fully connected	$56.83_{\pm 0.43}$	$83.56_{\pm 0.19}$
		Sparse	$\underline{57.18}_{\pm 0.19}$	$\underline{83.64}_{\pm 0.21}$

Table 8: Impact of graph topology on different convolutional operators. Results are reported on the classification task on the `CT-RATE` test set. The sparse topology is defined with receptive field size $q = 16$. Graph Convolution and Graph Attention operators are implemented with GraphConv and GATv2 from PyTorch Geometrics. For each operator, best results are underlined.

5.3 Robustness Analysis

Beyond overall classification accuracy, a clinically deployable model must remain reliable under common sources of variation in CT acquisitions. To this end, we assess the robustness of our approach to two perturbation types: *z-axis translation*, simulating patient body translation along the caudal–cranial axis (Di Piazza et al., 2025b), and *robustness to noise*, simulating variations in scanner calibration or patient-specific attenuation (Kiechle et al., 2024). Given the large number of baseline methods, we report robustness comparisons against the strongest representative of each family, as identified in our quantitative evaluations. This strategy eases understability and facilitates clearer interpretation.

Sensitivity to patient body translation. To emulate variability in patient positioning along the caudal–cranial axis, we translate the input volume by 5 to 30 axial slices in both directions, applying minimum-value padding to preserve dimensional consistency. Across all methods, Figure 9 shows that macro-F1 remains unchanged, indicating that both CT-SSG and the 3D-modeling baselines are robust to moderate volumetric misalignments.

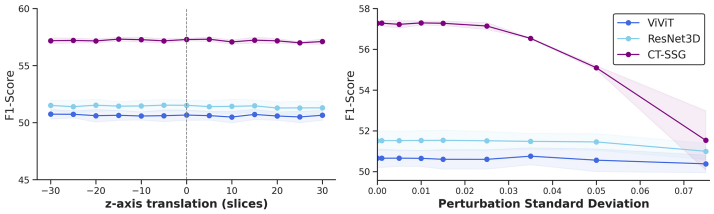

Figure 9: Robustness evaluation. *Left*: macro-F1 under axial z-axis translations (-30 to $+30$ slices), where all methods remain invariant to volumetric shifts. *Right*: macro-F1 under Gaussian noise perturbations of increasing standard deviation, where performance is stable up to $\sigma = 0.025$ and CT-SSG maintains higher F1-Score than baselines even as noise increases.

Sensitivity to noise. Following prior work (Sudre et al., 2017; Kiechle et al., 2024), we simulate acquisition-related variations by injecting Gaussian noise into voxel intensities, with standard deviation ranging from $\sigma_{min} = 0.01$ corresponding to low noise, to $\sigma_{max} = 0.07$ corresponding to high noise. While CT-SSG exhibits a mild decrease in performance at higher noise levels, it consistently maintains an advantage over the supervised baseline. This pattern suggests that pretraining confers robustness to common intensity variations.

5.4 Transfer on the Automated Report Generation task

Motivation. We further evaluate the visual encoders on an automated report generation task. To ensure that differences in downstream performance reflect the quality of the learned visual representations rather than decoder engineering, we adopt a deliberately simple encoder-decoder architecture inspired by CT2Rep (Hamamci et al., 2024). Concretely, the visual encoder is pretrained on the multi-label abnormality classification task and kept frozen while a lightweight decoder is trained with a next-token prediction objective. This setup isolates the quality of the latent space, which is the central focus of our study, from the confounding effects or more sophisticated sequence modeling strategies. Integration of advanced components such as large pretrained language models (Li et al., 2025), extensive modality-specific pretraining (Blankemeier et al., 2024) or multimodal fusion (Liu et al., 2025) if left for future work.

Evaluation protocol. Figure 11 provides an overview of the encoder-decoder pipeline. The pretrained visual encoder is frozen and a lightweight decoder is trained using a token-level cross-entropy loss for next-token prediction (Karpathy

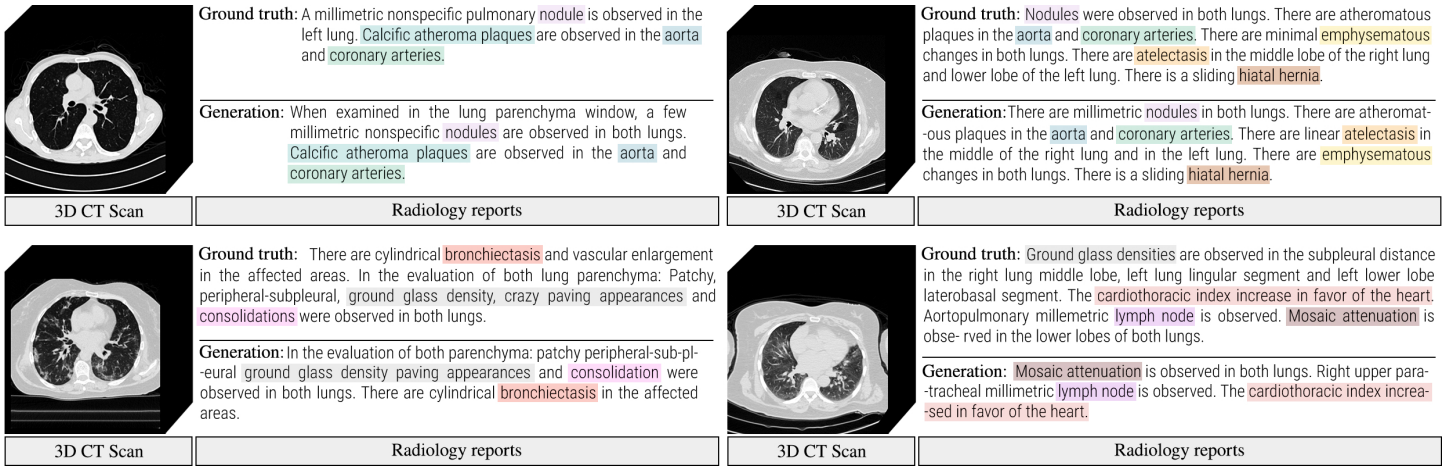

Figure 10: Qualitative comparison between ground-truth reports and those generated with CT-SSG for 3D Chest CT volumes. Color-coded highlights indicate abnormalities correctly captured by the model, demonstrating alignment with ground-truth annotations.

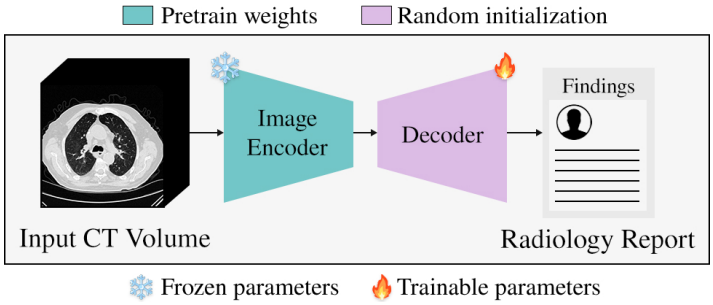

Figure 11: Report generation framework overview. The frozen pretrained image encoder extract visual features that are given to a decoder which generates the report, in a auto-regressively manner.

and Fei-Fei, 2015). At inference, reports are generated auto-regressively (Vinyals et al., 2015). This protocol ensures that observed differences in generated reports are primarily attributable to differences in the visual latent representations.

Quantitative results. We evaluate models using both Natural Language Generation (NLG) metrics and Clinical Efficacy (CE) metrics. NLG metrics, including BLEU-1 (Papineni et al., 2002) and METEOR (Lavie and Denkowski, 2009), assess the semantic alignment between ground-truth and generated reports. To assess clinical relevance, CE metrics quantify the model's ability to accurately identify and report pathologies present in the 3D CT Scans. Specifically, generated reports are given to a RadBERT (Yan et al., 2022) to extract predicted abnormalities as binary label vectors. These are compared against ground-truth annotations, enabling computation of standard classification metrics. We report F1-score and also incorporate the CRG Score, a recently proposed metric that weights positive predictions based on class frequency, reflecting the clinical imperative of minimizing false negatives, which can have serious consequences in medical diagnosis (Hamamci et al., 2025). Table 9 shows that our proposed CT-SSG achieves substantial improvements over baseline encoders on the report generation task. CT-SSG yields a relative gain of $+\Delta 40.56\%$ in F1-Score compared to ViViT, and $+\Delta 52.14\%$ compared to CNN-based baseline. A paired t-test between CT-SSG and each baseline results in p-values below 0.01 across all metrics, confirming the statistical significance of these improvements. Finally, Table 11 reports all metrics.

	NLG		CE	
Encoder	BLEU-1	METEOR	F1-Score	CRG
ViViT	0.290±0.008	0.150±0.004	27.54±0.40	40.01±0.22
ResNet3D	0.286±0.012	0.148±0.004	25.45±0.90	39.87±0.27
CT-SSG	<u>0.317</u>±0.001	<u>0.164</u>±0.003	<u>38.72</u>±0.25	<u>43.66</u>±0.15

Table 9: Quantitative evaluation on the report generation task, reporting both Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics on the `CT-RATE` dataset. CT-SSG achieves consistent improvements over baselines across metrics, with the <u>best results</u> underlined, demonstrating its ability to capture structured 3D information that benefits downstream report generation.

Qualitative results. Figure 10 presents qualitative examples comparing ground-truths with the generated reports. CT-SSG consistently identifies clinically relevant abnormalities and employs appropriate medical terminology, producing outputs that closely resemble radiologist-authored reports. Both quantitative and qualitative evaluations confirm that the model captures the presence of key abnormalities. However, descriptions of spatial localization and severity remain less reliable. Although automated report generation is not the primary focus of this study, these results underscore

the robustness of the learned representations and suggest a promising avenue for future research in connecting abnormality representation learning with clinically faithful language generation.

5.5 Transfer to Abdominal CT Scans for Multi-label Abnormality Classification

Motivation. We further investigate the *cross-domain generalization* of our learned representations by extending evaluation to a distinct anatomical region: 3D abdominal CT scans. Figure 12 suggests that this setting presents a substantial domain shift, as abdominal CTs capture a different anatomical field of view and spatial context compared to chest CTs.

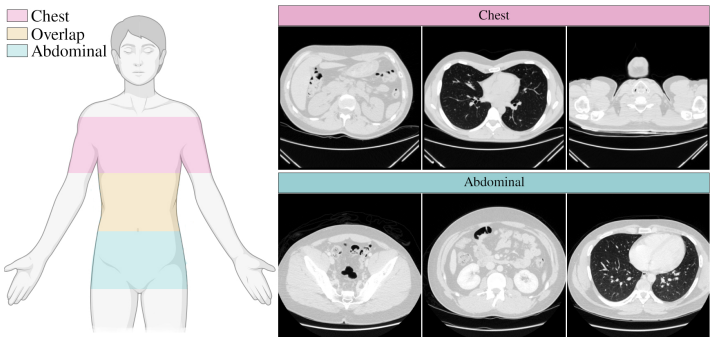

Figure 12: Comparison of chest and abdominal CT volumes. Top row: representative axial slices (first, center, last) from a chest CT volume, spanning from caudal to cranial directions (left to right). Bottom row: corresponding slices from an abdominal CT volume processed with the same spatial dimensions.

Evaluation protocol. To assess cross-anatomy generalization, we evaluate CT-SSG chest-pretrained model on the `Merlin Abdominal CT` dataset (Blankemeier et al., 2024). We focus on the 17 labels shared with `CT-RATE`, effectively probing whether features learned from chest CTs can transfer to detect the same abnormalities in a partially overlapping abdominal field of view. Expert annotations are not available, so we derive binary pseudo-labels from radiology reports using large language model-based inference (Reichenpfader et al., 2025). While these labels may introduce some noise, this approach provides a scalable way to test whether chest-pretrained representations encode medical priors that generalize across anatomical domains and imaging contexts.

Following established evaluation transfer protocols (Misra and Maaten, 2019; Bardes et al., 2021), we compare two configurations: a *supervised baseline* in which CT-SSG is trained from scratch with ImageNet-initialized ResNet weights, and a *linear probing*, in which a linear classifier is trained on top of frozen representations from our chest-pretrained CT-SSG backbone. Figure 13 illustrates the Linear probe framework. We vary the training set size between 250 and $10,000$ samples, with validation and test sets each comprising $1,000$ unique patients, ensuring subject-level separation.

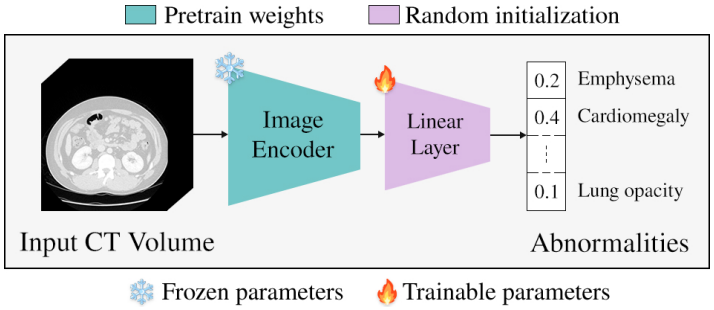

Figure 13: Linear probe framework overview. The frozen pretrained image encoder extract visual features that are given to a linear layer to predict abnormalities.

Quantitative results. Figure 14 reports macro-F1 and macro-mAP across training set sizes. In the low-data regime (100-3,750 samples), the *linear probe* consistently outperforms the supervised baseline, demonstrating that chest-pretrained representations encode transferable structural and textural priors that enable sample-efficient adaptation to abdominal CT. At larger training sizes, the fully supervised baseline surpasses the linear probe, as the increased availability of labeled data allows the model to specialize to the abdominal domain. These findings underscore the practical value of pretraining for clinically realistic, label-scarce scenarios.

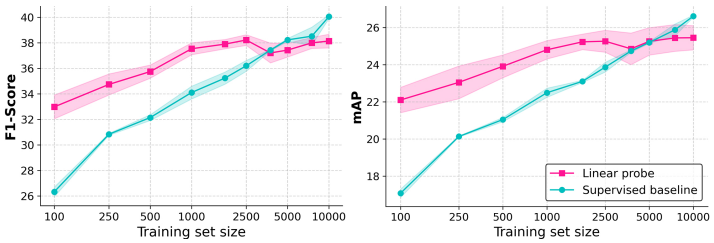

Figure 14: Transfer to abdominal CT. Comparison of a linear probe trained on frozen chest-pretrained CT-SSG representations against a supervised CT-SSG trained from scratch on the `Merlin Abdominal CT` dataset (Blankemeier et al., 2024). Performance is reported in terms of macro-F1 and macro-mAP for different size of the train set.

5.6 Implementation Details

Multi-label Abnormality Classification. CT-SSG was trained using the Adam optimizer (Kingma and Ba, 2017) with $(\beta_1, \beta_2) = (0.9, 0.99)$ and a learning rate of 0.0001. We used a batch size of 4, with $10,000$ warm-up steps

and $200,000$ iterations to ensure convergence. The training duration of CT-SSG was one day on a single NVIDIA RTX A6000 GPU. We used a GPU with 48GB of memory but the training can only require 16GB of memory with implementation of gradient accumulation. Inference takes approximately 70 milliseconds.

Report Generation. The Encoder-Decoder report generation framework from Section 5.4, was trained using Adam with a batch size of 4, $(\beta_1, \beta_2) = (0.9, 0.99)$, and a learning rate of 0.00005 for $400,000$ iterations to ensure convergence. At inference, we use the Beam Search algorithm as generation mode (Freitag and Al-Onaizan, 2017) with a beam size set to 4, and generated sentences were limited to 300 tokens. Inference takes approximately 0.90 seconds per generated report.

Linear probe. The linear probe on abdominal CTs, presented in Section 5.5, was trained using Adam with $(\beta_1, \beta_2) = (0.9, 0.99)$, a learning rate of 0.0001, and a batch size of 4 for $200,000$ iterations.

6. Conclusion and Discussion

In this work, we introduced CT-SSG, a 2.5D approach that represents 3D CT Volumes as structured graphs constructed from axial slices. Evaluated on chest CT datasets for multi-label abnormality classification, CT-SSG achieves competitive performance while maintaining computational efficiency compatible with clinical deployment. Specifically, restricting inter-slice connectivity to local neighborhoods, rather than adopting a fully-connected transformer-style topology, yields higher clinical accuracy, suggesting that explicit structural priors can serve as an effective inductive bias for 3D reasoning. Our ablation studies confirm that graph topology, positional encoding, and aggregation operators play complementary roles in enabling this sparse yet expressive representation. Beyond classification, we demonstrated the transferability of the learned representations to radiology report generation and abdominal CT, highlighting the robustness and generality of the proposed approach.

Limitations and Future work. (1) Our work primarily focuses on graph-based feature aggregation and employs a 2D ResNet-18 encoder for slice-level feature initialization. While this choice allows us to isolate the contribution of the proposed aggregation mechanism, the CT-SSG formulation remains encoder-agnostic. Future research could investigate the integration of higher-capacity visual encoders and advanced pretraining strategies, which may further improve representation learning and overall performance. (2) Since CT-SSG relies on spectral convolutions, the learned filters are linked to the underlying graph topology. Although the proposed graph construction yields highly consistent structures across volumes due to spatial standardization and anatomical ordering, substantial variations in acquisition protocols, field-of-view, or graph construction strategies may require adaptation or retraining. Future work could investigate topology-robust graph learning or domain adaptation strategies to improve generalization across heterogeneous imaging settings. (3) As a 2.5D, axial-slice–based method, CT-SSG does not fully exploit volumetric information. A promising direction is to hybridize with a fully 3D branch or adopt multi-view representations that integrate sagittal and coronal planes to complement axial features. (4) In transferring to automated report generation, CT-SSG consistently identifies key abnormalities but is less reliable in describing their spatial localization and severity. As report generation was not the primary objective of this study, these findings underscore the versatility of the learned representations while pointing toward an important future direction: bridging abnormality representation learning with clinically faithful narrative generation.

Acknowledgments

We acknowledge `CT-RATE`, `RAD-ChestCT`, and `Merlin CT` authors for releasing their public datasets to be used for this work of academic research. The patient body icon from Figure 12 was created with *BioRender.com*. Finally, we thank the anonymous reviewers and editors of the MICCAI EMERGE Workshop and of the MELBA Journal for their valuable feedback and suggestions.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

The authors declare no conflict of interest.

Data availability

The `CT-RATE` public dataset is available at `https://huggingface.co/datasets/ibrahimhamamci/CT-RATE`. The `RAD-ChestCT` public dataset is available at `https://zenodo.org/records/6406114`. The `Merlin Abdominal CT` public dataset is available at `https://stanfordaimi.azurewebsites.net/datasets/60b9c7ff-877b-48ce-96c3-0194c8205c40`.

References

David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future. *Sensors (Basel, Switzerland)*, 21(14):4758, July 2021. ISSN 1424-8220. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8309939/.

Andrés Anaya-Isaza, Leonel Mera-Jiménez, and Martha Zequera-Diaz. An overview of deep learning in medical imaging. *Informatics in Medicine Unlocked*, 26:100723, January 2021. ISSN 2352-9148. . URL https://www.sciencedirect.com/science/article/pii/S2352914821002033.

Prasanna Sakthi Aravazhi, Praveen Gunasekaran, Neo Zhong Yi Benjamin, Andy Thai, Kiran Kishor Chandrasekar, Nikhil Deep Kolanu, Priyadarshi Prajjwal, Yogesh Tekuru, Lissette Villacreses Brito, and Pugazhendi Inban. The integration of artificial intelligence into clinical medicine: Trends, challenges, and future directions. *Disease-a-month: DM*, 71(6):101882, June 2025. ISSN 1557-8194. .

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-6654-2812-5. . URL https://ieeexplore.ieee.org/document/9710415/.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization, July 2016. URL http://arxiv.org/abs/1607.06450. arXiv:1607.06450 [cs, stat].

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate, May 2016. URL http://arxiv.org/abs/1409.0473. arXiv:1409.0473 [cs].

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. October 2021. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Maya Bechler-Speicher, Amir Globerson, and Ran Gilad-Bachrach. The Intelligible and Effective Graph Neural Additive Networks, December 2024. URL http://arxiv.org/abs/2406.01317. arXiv:2406.01317 [cs].

Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/hash/f106b7f99d2cb30c3db1c3cc0fde9ccb-Abstract.html.

Mollie Bianchi and Timothy D. Barfoot. UAV Localization Using Autoencoded Satellite Images, February 2021. URL http://arxiv.org/abs/2102.05692. arXiv:2102.05692 [cs].

Louis Blankemeier, Joseph Paul Cohen, Ashwin Kumar, Dave Van Veen, Syed Jamal Safdar Gardezi, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Cesar Truyts, Christian Bluethgen, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: A Vision Language Foundation Model for 3D Computed Tomography, June 2024. URL http://arxiv.org/abs/2406.06512. arXiv:2406.06512 [cs] version: 1.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan

Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

Joshua Broder and David M. Warshauer. Increasing utilization of computed tomography in the adult emergency department, 2000-2005. *Emergency Radiology*, 13(1): 25–30, October 2006. ISSN 1070-3004. .

Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks?, January 2022. URL http://arxiv.org/abs/2105.14491. arXiv:2105.14491 [cs].

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs, May 2014. URL http://arxiv.org/abs/1312.6203. arXiv:1312.6203 [cs].

Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, February 2018. URL http://arxiv.org/abs/1705.07750. arXiv:1705.07750 [cs].

Hao-Hsiang Chang, Yu-Hua Chang, Yi-Lung Shih, Cheng-Hsun Lin, and Huang-Chia Shih. Basketball Player Action Recognition and Tracking Using R(2+1)D CNN With Spatial-temporal Features. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 388–389, October 2024. . URL https://ieeexplore.ieee.org/document/10760677.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering, February 2017. URL http://arxiv.org/abs/1606.09375. arXiv:1606.09375 [cs].

Tami D. DenOtter and Johanna Schubert. Hounsfield Unit. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL http://www.ncbi.nlm.nih.gov/books/NBK547721/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, October 2018. URL https://arxiv.org/abs/1810.04805v2.

Theo Di Piazza, Carole Lazarus, Olivier Nempont, and Loic Boussel. Imitating Radiological Scrolling: A Global-Local Attention Model for 3D Chest CT Volumes Multi-Label Anomaly Classification. *MIDL*, January 2025a. URL https://openreview.net/forum?id=YPBbNAmGXW.

Theo Di Piazza, Carole Lazarus, Olivier Nempont, and Loic Boussel. Structured Spectral Graph Learning for Anomaly Classification in 3D Chest CT Scans, August 2025b. URL http://arxiv.org/abs/2508.01045. arXiv:2508.01045 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Rachel Lea Draelos, David Dov, Maciej A. Mazurowski, Joseph Y. Lo, Ricardo Henao, Geoffrey D. Rubin, and Lawrence Carin. Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Medical Image Analysis*, 67: 101857, January 2021. ISSN 1361-8415. . URL https://www.sciencedirect.com/science/article/pii/S1361841520302218.

Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric, April 2019. URL http://arxiv.org/abs/1903.02428. arXiv:1903.02428 [cs].

Markus Freitag and Yaser Al-Onaizan. Beam Search Strategies for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, 2017. . URL http://arxiv.org/abs/1702.01806. arXiv:1702.01806 [cs].

Francesco Di Giovanni, Lorenzo Giusti, Federico Barbero, Giulia Luise, Pietro Lio', and Michael Bronstein. On Over-Squashing in Message Passing Neural Networks: The Impact of Width, Depth, and Topology, May 2023. URL http://arxiv.org/abs/2302.02941. arXiv:2302.02941 [cs].

Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. HIGT: Hierarchical Interaction Graph-Transformer for Whole Slide Image Analysis, September 2023. URL http://arxiv.org/abs/2309.07400. arXiv:2309.07400 [cs].

Ibrahim Ethem Hamamci, Sezgin Er, Enis Simsar, Anjany Sekuboyina, Chinmay Prabhakar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, Hadrien Reynaud, Sarthak Pati, Christian Bluethgen, Mehmet Kemal Ozdemir, and Bjoern Menze. GenerateCT: Text-Conditional Generation of 3D Chest CT Volumes, November 2023. URL http://arxiv.org/abs/2305.16037. arXiv:2305.16037 [cs].

Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern Menze. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. March 2024. URL http://arxiv.org/abs/2403.06801. arXiv:2403.06801 [cs, eess].

Ibrahim Ethem Hamamci, Sezgin Er, Suprosanna Shit, Hadrien Reynaud, Bernhard Kainz, and Bjoern Menze. CRG Score: A Distribution-Aware Clinical Metric for Radiology Report Generation, May 2025. URL http://arxiv.org/abs/2505.17167. arXiv:2505.17167 [cs].

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, Weicheng Dai, Murong Xu, Hadrien Reynaud, Muhammed Furkan Dasdelen, Bastian Wittmann, Tamaz Amiranashvili, Enis Simsar, Mehmet Simsar, Emine Bensu Erdemir, Abdullah Alanbay, Anjany Sekuboyina, Berkan Lafci, Ahmet Kaplan, Zhiyong Lu, Malgorzata Polacin, Bernhard Kainz, Christian Bluethgen, Kayhan Batmanghelich, Mehmet Kemal Ozdemir, and Bjoern Menze. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nature Biomedical Engineering*, February 2026. ISSN 2157-846X. . URL https://www.nature.com/articles/s41551-025-01599-y.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

Ademola E. Ilesanmi, Taiwo O. Ilesanmi, and Babatunde O. Ajayi. Reviewing 3D convolutional neural network approaches for medical image segmentation. *Heliyon*, 10 (6):e27398, March 2024. ISSN 2405-8440. . URL https://www.sciencedirect.com/science/article/pii/S2405844024034297.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, January 2019. URL http://arxiv.org/abs/1901.07031. arXiv:1901.07031 [cs, eess].

Alistair E. W. Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. The MIMIC-CXR Database, 2019. URL https://physionet.org/content/mimic-cxr/.

Chaitanya K. Joshi. Transformers are Graph Neural Networks, June 2025. URL http://arxiv.org/abs/2506.22084. arXiv:2506.22084 [cs].

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873):583–589, August 2021. ISSN 1476-4687. . URL https://www.nature.com/articles/s41586-021-03819-2.

Emilien Jupin-Delevaux, Aissam Djahnine, François Talbot, Antoine Richard, Sylvain Gouttard, Adeline Mansuy, Philippe Douek, Salim Si-Mohamed, and Loïc Boussel. BERT-based natural language processing analysis of French CT reports: Application to the measurement of the positivity rate for pulmonary embolism. *Research in Diagnostic and Interventional Imaging*, 6:100027, June 2023. ISSN 2772-6525. .

Hamza Kalisch, Fabian Hörst, Jens Kleesiek, Ken Herrmann, and Constantin Seibold. CT-GRAPH: Hierarchical Graph Attention Network for Anatomy-Guided CT Report Generation, August 2025. URL http://arxiv.org/abs/2508.05375. arXiv:2508.05375 [cs].

Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions, April 2015. URL http://arxiv.org/abs/1412.2306. arXiv:1412.2306 [cs].

Johannes Kiechle, Daniel M. Lang, Stefan M. Fischer, Lina Felsner, Jan C. Peeken, and Julia A. Schnabel. Graph Neural Networks: A suitable Alternative to MLPs in Latent 3D Medical Image Classification?, July 2024. URL http://arxiv.org/abs/2407.17219. arXiv:2407.17219 [cs].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. A Survey on Biomedical Image Captioning, May 2019. URL http://arxiv.org/abs/1905.13302. arXiv:1905.13302 [cs] version: 1.

Alon Lavie and Michael J. Denkowski. The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2):105–115, September 2009. ISSN 1573-0573. . URL https://doi.org/10.1007/s10590-009-9059-4.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. ISSN 0028-0836, 1476-4687. . URL https://www.nature.com/articles/nature14539.

Siyou Li, Pengyao Qin, Huanan Wu, Dong Nie, Arun J Thirunavukarasu, Juntao Yu, and Le Zhang. μ2 Tokenizer: Differentiable Multi-Scale Multi-Modal Tokenizer for Radiology Report Generation. 2025.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation, June 2021a. URL http://arxiv.org/abs/2106.06963. arXiv:2106.06963 [cs].

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced Contrastive Learning with Multi-view Longitudinal Data for Chest X-ray Report Generation, February 2025. URL http://arxiv.org/abs/2502.20056. arXiv:2502.20056 [cs].

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021b. URL http://arxiv.org/abs/2103.14030. arXiv:2103.14030 [cs].

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer, June 2021c. URL http://arxiv.org/abs/2106.13230. arXiv:2106.13230 [cs].

Jun Ma, Feifei Li, and Bo Wang. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation, January 2024. URL http://arxiv.org/abs/2401.04722. arXiv:2401.04722 [eess].

Nikita Makarov, Santhanakrishnan Narayanan, and Constantinos Antoniou. Graph neural network surrogate for strategic transport planning, August 2024. URL http://arxiv.org/abs/2408.07726. arXiv:2408.07726 [cs].

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-Entropy Loss Functions: Theoretical Analysis and Applications, June 2023. URL http://arxiv.org/abs/2304.07288. arXiv:2304.07288 [cs].

Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations, December 2019. URL http://arxiv.org/abs/1912.01991. arXiv:1912.01991 [cs].

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman Go Neural: Higher-order Graph Neural Networks, November 2021. URL http://arxiv.org/abs/1810.02244. arXiv:1810.02244 [cs].

Gustav Müller-Franzes, Firas Khader, Robert Siepmann, Tianyu Han, Jakob Nikolas Kather, Sven Nebelung, and Daniel Truhn. Medical slice transformer for improved diagnosis and explainability on 3D medical images with DINOv2. *Scientific Reports*, 15(1):23979, July 2025. ISSN 2045-2322. . URL https://www.nature.com/articles/s41598-025-09041-8.

Reabal Najjar. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. *Diagnostics*, 13(17):2760, August 2023. ISSN 2075-4418. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10487271/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Suraj Pai, Ibrahim Hadzic, Dennis Bontempi, Keno Bressem, Benjamin H. Kann, Andriy Fedorov, Raymond H. Mak, and Hugo J. W. L. Aerts. Vision Foundation Models for Computed Tomography, February 2025. URL http://arxiv.org/abs/2501.09001. arXiv:2501.09001 [eess].

Harsh Panwar, P.K. Gupta, Mohammad Khubeb Siddiqui, Ruben Morales-Menendez, Prakhar Bhardwaj, and Vaishnavi Singh. A deep learning and grad-CAM based color visualization approach for fast detection of COVID-19 cases using chest X-ray and CT-Scan images. *Chaos, Solitons, and Fractals*, 140:110190, November 2020. ISSN 0960-0779. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7413068/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. . URL https://aclanthology.org/P02-1040.

Paula R. Patel and Orlando De Jesus. CT Scan. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL http://www.ncbi.nlm.nih.gov/books/NBK567796/.

Daniel Reichenpfader, Henning Muller, and Kerstin Denecke. A scoping review of large language model based approaches for information extraction from radiology reports | npj Digital Medicine, 2025. URL https://www.nature.com/articles/s41746-024-01219-0.

Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):1–18, November 2022. ISSN 2662-4443. . URL https://www.nature.com/articles/s43246-022-00315-6.

Amanda Ross and Victor L. Willson. Paired Samples T-Test. In Amanda Ross and Victor L. Willson, editors, *Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures*, pages 17–19. SensePublishers, Rotterdam, 2017. ISBN 978-94-6351-086-8. . URL https://doi.org/10.1007/978-94-6351-086-8_4.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. URL http://arxiv.org/abs/1409.0575. arXiv:1409.0575 [cs].

Daniel Scholz, Ayhan Can Erdur, Viktoria Ehm, Anke Meyer-Baese, Jan C. Peeken, Daniel Rueckert, and Benedikt Wiestler. MM-DINOv2: Adapting Foundation Models for Multi-Modal Medical Image Analysis, September 2025. URL http://arxiv.org/abs/2509.06617. arXiv:2509.06617 [eess].

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, December 2019. URL http://arxiv.org/abs/1610.02391. arXiv:1610.02391.

Noam Shazeer. GLU Variants Improve Transformer, February 2020. URL http://arxiv.org/abs/2002.05202. arXiv:2002.05202 [cs].

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. ISSN 0035-9246. URL https://www.jstor.org/stable/2984809.

Carole H. Sudre, M. Jorge Cardoso, Sebastien Ourselin, and Alzheimer's Disease Neuroimaging Initiative. Longitudinal segmentation of age-related white matter hyperintensities. *Medical Image Analysis*, 38:50–64, May 2017. ISSN 1361-8423. .

Yucheng Tang, Dong Yang, Wenqi Li, Holger Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis, March 2022. URL http://arxiv.org/abs/2111.14791. arXiv:2111.14791 [cs].

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and Explainable Region-guided Radiology Report Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7433–7442, June 2023. . URL http://arxiv.org/abs/2304.08295. arXiv:2304.08295 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Petar Veličković. Everything is Connected: Graph Neural Networks. *Current Opinion in Structural Biology*, 79:102538, April 2023. ISSN 0959440X. . URL http://arxiv.org/abs/2301.08210. arXiv:2301.08210 [cs].

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, February 2018. URL http://arxiv.org/abs/1710.10903. arXiv:1710.10903 [stat].

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator, April 2015. URL http://arxiv.org/abs/1411.4555. arXiv:1411.4555 [cs].

Tassilo Wald, Ibrahim Ethem Hamamci, Yuan Gao, Sam Bond-Taylor, Harshita Sharma, Maximilian Ilse, Cynthia Lo, Olesya Melnichenko, Anton Schwaighofer, Noel C. F. Codella, Maria Teodora Wetscherek, Klaus H. Maier-Hein, Panagiotis Korfiatis, Valentina Salvatelli, Javier Alvarez-Valle, and Fernando Pérez-García. Comprehensive language-image pre-training for 3D medical image understanding, January 2026. URL http://arxiv.org/abs/2510.15042. arXiv:2510.15042 [cs].

Chuqi Wang. A Review on 3D Convolutional Neural Network. In *2023 IEEE 3rd International Conference on*

Power, Electronics and Computer Applications (ICPECA), pages 1204–1208, January 2023. . URL https://ieeexplore.ieee.org/document/10075760.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, July 2022. . URL https://pubs.rsna.org/doi/full/10.1148/ryai.210258.

Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding, August 2023. URL http://arxiv.org/abs/2304.06906. arXiv:2304.06906 [cs].

Xitong Zhang, Yixuan He, Nathan Brugnone, Michael Perlmutter, and Matthew Hirn. MagNet: A Neural Network for Directed Graphs, June 2021. URL http://arxiv.org/abs/2102.11391. arXiv:2102.11391 [cs].

Yuhui Zhang, Shih-Cheng Huang, Zhengping Zhou, Matthew P. Lungren, and Serena Yeung. Adapting Pretrained Vision Transformers from 2D to 3D through Weight Inflation Improves Medical Image Segmentation, February 2023. URL http://arxiv.org/abs/2302.04303. arXiv:2302.04303 [cs].

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, January 2020. ISSN 2666-6510. . URL https://www.sciencedirect.com/science/article/pii/S2666651021000012.

Appendix A. Per-abnormality F1-Score

Table 10 presents a per-abnormality performance comparison between CT-SSG and competing baseline methods. All models are trained and evaluated on the CT-RATE dataset, and the reported results correspond to the mean performance across a 5-fold cross-validation.

Appendix B. Model capacity and performance

To verify that the observed performance improvements are not merely due to increased model capacity, Figure 15 displays the F1-Score, AUROC and mAP against the number of learnable parameters. Our method achieves superior performance with a comparable parameter count, indicating that the gains arise from the proposed spectral representation rather than model scaling.

Appendix C. Pseudo code

Table 13 presents a step-by-step pseudo code for CT-SSG implementation.

Appendix D. t-SNE visualization

Figure 16 illustrates a t-SNE visualization of the pooled features vector, denoted as $\bar{z}$ and defined in Equation 10. The t-SNE was implemented using scikit-learn t-SNE module, with a 2-dimensional latent space and a perplexity of 30.

Appendix E. Comparison with Generalist Computed Tomography Foundation Models

Table 14 summarizes the configurations of the evaluated CT-specific foundation models under the linear probing setting. For each model, latent representations are extracted from pretrained and frozen visual encoders. Preprocessing follows the protocols reported in the corresponding original works, including volume orientation, spatial resolution, and Hounsfield Unit (HU) range, to ensure a fair comparison. To further contextualize multi-label abnormality classification performance, the table also reports pretraining objectives, training datasets and their sizes, anatomical coverage, and the latent dimensionality of the backbone representations.

Appendix F. Training protocols

Table 15 summaries backbone initialization and training protocols for all methods. Table 16 provides links to the baseline implementations and model weights for initialization or direct comparison.

Appendix G. Effect of Slice Overlap in 2.5D Graph Construction

To analyze the effect of overlapping versus non-overlapping axial slice triplets in the proposed framework, Table 17 reports the number of graph nodes, training and inference times, and multi-label abnormality classification performance for both configurations. The two settings achieve comparable classification performance, indicating that introducing overlap provides no measurable performance benefit. In contrast, the overlapping configuration substantially increases the number of graph nodes as well as the training time to convergence and inference time. Based on this performance-efficiency trade-off, the non-overlapping configuration was adopted in the main experiments.

Appendix H. 2D Backbone scaling

While our work primarily focuses on graph-based feature aggregation and deliberately adopts ResNet-18 as an efficient 2.5D backbone, we evaluate the scalability of CT-SSG using a higher-capacity encoder. To ensure a fair comparison, we replace the ResNet-18 backbone with ResNet-34 while preserving all other architectural and training settings. Table 18 reports parameter counts and performance obtained with both encoders. The results demonstrate that CT-SSG benefits from increased backbone capacity, highlighting the backbone-agnostic design of CT-SSG and its ability to scale with stronger feature extractors.

Appendix I. Number of axial slices per node

CT-SSG design models the CT Scan as a graph of nodes that correspond to triplets of axial slices, allowing to capture inter-slice anatomical continuity and subtle volumetric variations while preserving high in-plane resolution by matching the 3 channel RGB structure of 2D backbones for initial feature extraction. To analyze how the number of axial slices per node impact performance, we further investigate CT-SSG performance with 1 unique axial slice per node and 6 unique slices per node. To ensure fair comparison across configurations, we conserve the same CT-SSG's architectural components and extract initial node features by inflating the 2D ResNet backbone weights into 3D, following Carreira and Zisserman (2018), a strategy that enables to vary the number of axial slices encoded within a node. Table 19 reports multi-label abnormality classification performance for 1, 3 and 6 adjacent axial slices per node. CT-SSG yields the best results with 3 axial slices per node, whose initial features are extracted by a 2D ResNet. We observe that CT-SSG yields the best performance with 3 axial slices per node, achieving a favourable balance between local anatomical continuity and representations compactness.

Abnormality	ViT3D	ViViT	Swin3D	ResNet3D	CT-Net	CT-MvG	CT-Scroll	CT-SSG
Medical material	0.374	0.373	0.380	0.403	0.450	0.393	0.540	0.631
Arterial wall calcification	0.701	0.699	0.692	0.738	0.771	0.716	0.758	0.778
Cardiomegaly	0.551	0.554	0.608	0.564	0.586	0.574	0.587	0.624
Pericardial effusion	0.401	0.388	0.400	0.401	0.385	0.378	0.420	0.554
Coronary artery wall calcif.	0.647	0.655	0.628	0.674	0.760	0.656	0.768	0.767
Hiatal hernia	0.360	0.357	0.358	0.359	0.346	0.365	0.362	0.421
Lymphadenopathy	0.504	0.502	0.499	0.504	0.487	0.522	0.511	0.520
Emphysema	0.467	0.474	0.470	0.475	0.450	0.506	0.487	0.509
Atelectasis	0.454	0.481	0.469	0.476	0.466	0.503	0.494	0.521
Lung nodule	0.632	0.632	0.633	0.637	0.633	0.637	0.634	0.637
Lung opacity	0.638	0.681	0.676	0.701	0.716	0.703	0.738	0.738
Pulmonary fibrotic sequela	0.441	0.442	0.443	0.449	0.433	0.457	0.456	0.467
Pleural effusion	0.768	0.756	0.776	0.810	0.776	0.778	0.817	0.825
Mosaic attenuation	0.361	0.395	0.387	0.408	0.338	0.423	0.394	0.456
Peribronchial thickening	0.355	0.366	0.343	0.356	0.343	0.394	0.384	0.394
Consolidation	0.558	0.603	0.604	0.600	0.560	0.625	0.641	0.641
Bronchiectasis	0.292	0.347	0.304	0.333	0.304	0.373	0.367	0.391
Interlobular septal thick.	0.413	0.415	0.377	0.384	0.374	0.418	0.417	0.418
Mean	0.495	0.507	0.503	0.515	0.510	0.524	0.543	0.572

Table 10: Per-abnormality F1-Score of all methods evaluated on the CT-RATE test set, averaged across 5 folds. The underlined metrics are those that have improved with CT-SSG compared to baselines. *Calcif.* denotes for *calcification* and *thick.* for *thickening*. CT-SSG demonstrates improved F1-Scores across the majority of abnormalities.

	Natural Language Generation				Clinical Efficacy			
Encoder	**BLEU-1**	**BLEU-4**	**METEOR**	**ROUGE**	**F1-Score**	**Precision**	**Recall**	**Accuracy**
ResNet3D	0.286	0.106	0.148	0.242	0.254	0.384	0.236	0.816
ViViT	0.290	0.106	0.150	0.241	0.275	0.418	0.253	0.808
CT-SSG	0.317	0.107	0.164	0.246	0.387	0.520	0.363	0.814

Table 11: Detailed automated report generation metrics evaluated on the CT-RATE test set, averaged across 5 folds.

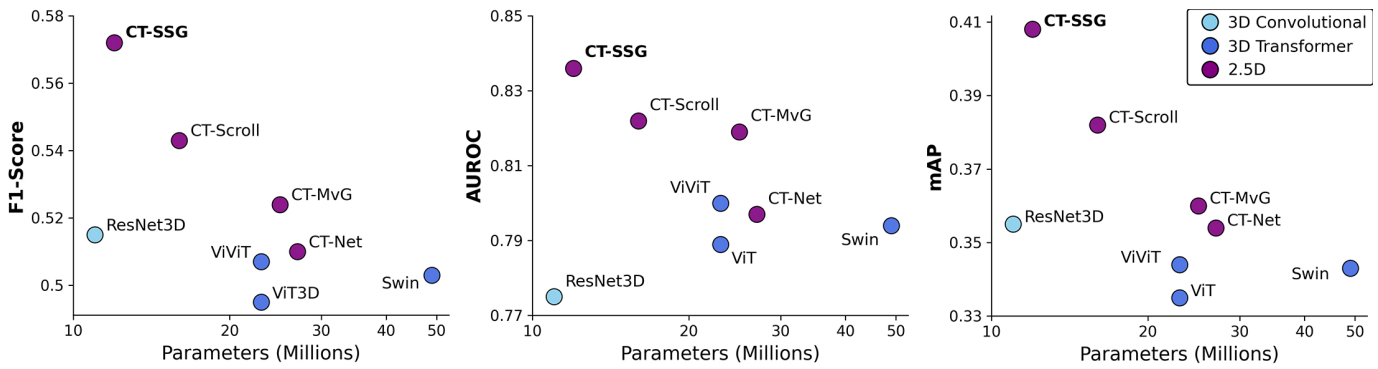

Figure 15: Comparison of F1-Score, AUROC and mAP across models with varying parameter counts. Despite comparable model capacities, CT-SSG achieves consistently higher performance, suggesting that the performance gains stem from improved representation learning rather than increased model size.

Abnormality	F1-Score	Precision	Recall	AUROC	AP	Accuracy
Medical material	0.631	0.561	0.674	0.905	0.647	0.914
Arterial wall calcification	0.778	0.729	0.814	0.928	0.812	0.865
Cardiomegaly	0.624	0.532	0.702	0.935	0.660	0.901
Pericardial effusion	0.554	0.499	0.625	0.902	0.586	0.928
Coronary artery wall calcification	0.767	0.680	0.866	0.931	0.782	0.869
Hiatal hernia	0.421	0.322	0.563	0.773	0.410	0.770
Lymphadenopathy	0.520	0.419	0.693	0.743	0.509	0.676
Emphysema	0.509	0.474	0.518	0.789	0.546	0.794
Atelectasis	0.521	0.441	0.675	0.776	0.543	0.726
Lung nodule	0.637	0.493	0.907	0.626	0.540	0.547
Lung opacity	0.738	0.735	0.742	0.866	0.824	0.801
Pulmonary fibrotic sequela	0.467	0.350	0.722	0.683	0.462	0.567
Pleural effusion	0.825	0.826	0.832	0.969	0.853	0.959
Mosaic attenuation pattern	0.456	0.498	0.410	0.884	0.488	0.919
Peribronchial thickening	0.394	0.324	0.554	0.798	0.399	0.808
Consolidation	0.641	0.575	0.721	0.901	0.662	0.846
Bronchiectasis	0.391	0.327	0.504	0.787	0.422	0.838
Interlobular septal thicking	0.418	0.369	0.446	0.860	0.365	0.892
Mean	0.572	0.509	0.665	0.836	0.582	0.812

Table 12: Per-abnormality classification metrics for CT-SSG, evaluated on the `CT-RATE` test set and averaged across 5 folds. We report macro F1-Score with Precision, Recall, AUROC, Average Precision (AP) and Accuracy.

Section	Pseudo code	Description	Input shape	Output shape	Python module
Features Initialization	$\mathcal{X} \leftarrow \text{LoadCTVolume}()$	Load 3D CT Volumes	–	$(B, 1, S, H_s, W_s)$	–
	$\mathcal{X}^{\text{triplet}} \leftarrow \text{MakeTriplet}(\mathcal{X})$	Group triplet of axial slices	$(B, 1, S, H_s, W_s)$	$(B, N, 3, H_s, W_s)$	`torch.reshape`[1]
	$H^{\text{resnet}} \leftarrow \text{ResNet}(\mathcal{X}^{\text{triplet}})$	2D ResNet forward pass	$(B, N, 3, H_s, W_s)$	(B, N, d, h, w)	`resnet18`[2]
	$\bar{H} \leftarrow \text{GAP}(H^{\text{resnet}})$	Global Average Pooling	(B, N, d, h, w)	(B, N, d)	`torch.mean`[1]
	$H \leftarrow \bar{H} + P^{\text{axial}}_{\text{pos}}$	Positional Embedding	(B, N, d)	(B, N, d)	–
	$H_0 \leftarrow H$	Features initialization	(B, N, d)	(B, N, d)	–
Message Passing ($\times L$)	$H_l^{'} \leftarrow \text{LayerNorm}(H_l)$	Layer Normalization	(B, N, d)	(B, N, d)	`nn.LayerNorm`[1]
	$E^{\text{index}} \leftarrow \text{LoadEdgeIndex}()$	Load edges index	–	$(2, N \times N)$	`torch.Tensor`[1]
	$E^{\text{weight}} \leftarrow \text{LoadEdgeWeight}()$	Load edges weight	–	$(N \times N)$	`torch.Tensor`[1]
	$\mathcal{G} \leftarrow \text{BuildGraph}(H_l^{'}, E^{\text{index}}, E^{\text{weight}})$	Build graph	–	–	`pyg.data.Data`[3]
	$H_l^{'} \leftarrow \text{ChebConv}(\mathcal{G})$	Chebyshev Convolution	(B, N, d)	(B, N, d)	`pyg.nn.conv.Chebconv`[3]
	$Z_l \leftarrow H_l^{'} + H_l$	Residual Connection	(B, N, d)	(B, N, d)	–
	$Z_l^{'} \leftarrow \text{LayerNorm}(Z_l)$	Layer Normalization	(B, N, d)	(B, N, d)	`nn.LayerNorm`[1]
	$Z_l^{'} \leftarrow \text{FFN}(Z_l^{'})$	Feedforward Neural Network	(B, N, d)	(B, N, d)	`nn.Linear, nn.GELU`[1]
	$H_{l+1} \leftarrow Z_l^{'} + Z_l$	Residual Connection	(B, N, d)	(B, N, d)	–
Classification	$\bar{z} \leftarrow \text{Pooling}(H_L)$	Graph pooling	(B, N, d)	(B, d)	`torch.mean`[1]
	$\hat{y} \leftarrow \text{Classifier}(\bar{z})$	Classification head	(B, d)	(B, M)	`torch.Linear`[1]

Table 13: Step-by-step pseudo code of CT-SSG with semantic description, tensor shapes, and python modules. [1] refers to `PyTorch` module. [2] refers to `torchvision` module. [3] refers to `PyTorch Geometric` module.

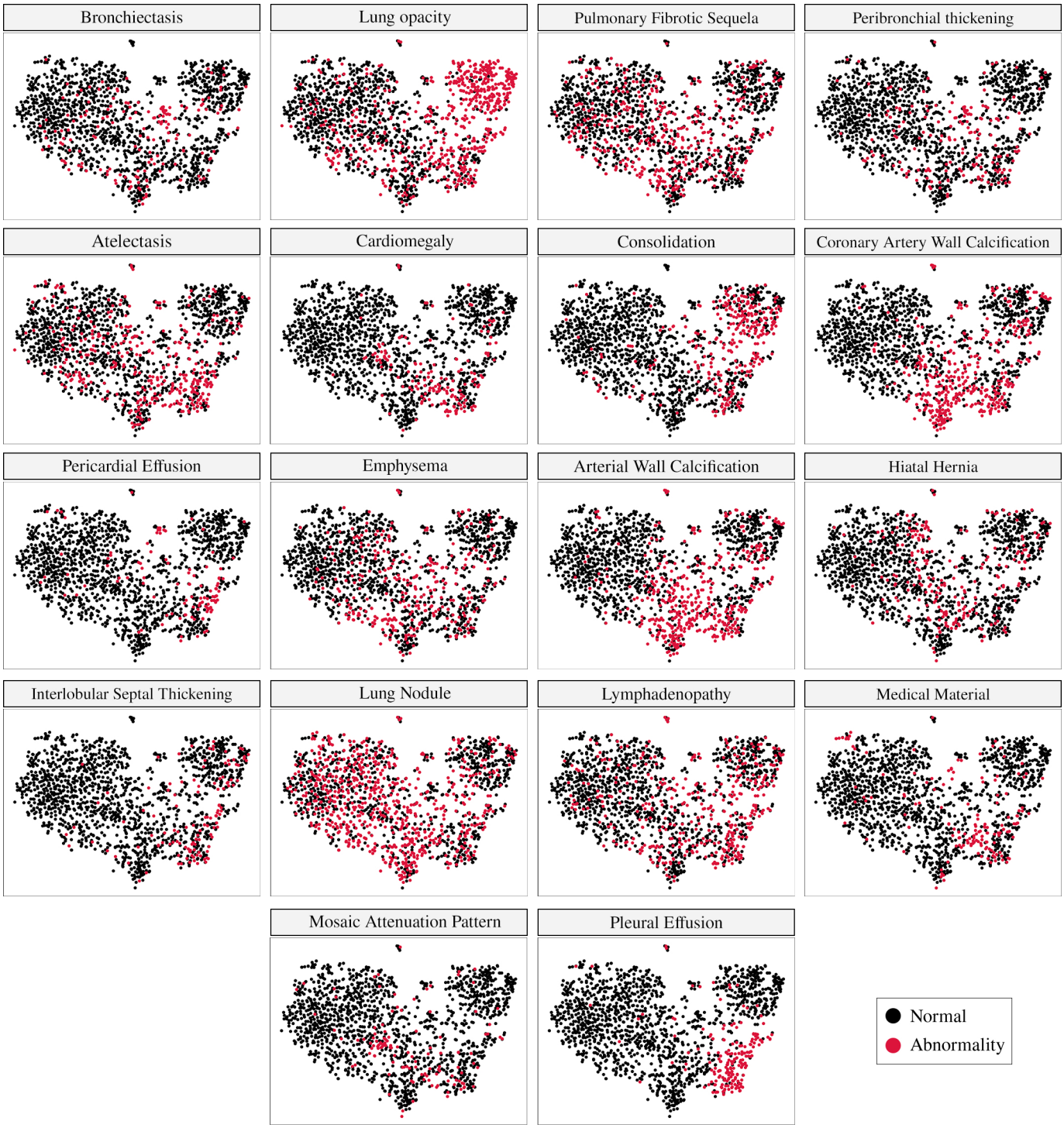

Figure 16: t-SNE visualization, of the pooled features for the `CT-RATE` test dataset. The colors represent classes.

Method			Data			Processing		Complexity		RAD-ChestCT Metrics	
Setup	Model	Pretrain task(s)	Dataset(s)	Train size	Region(s)	Spacing (mm)	HU range	Param.	Dim.	AUROC	mAP
Foundation Models Probing	**CT-FM**	Contrastive	69 cohorts[1]	148k	Chest, Abdomen[2]	$(3.0, 1.0, 1.0)$	$[-1024, 2048]$	77M	512	$63.26_{\pm 0.13}$	$43.78_{\pm 0.15}$
	CT-CLIP	Vision-language	CT-RATE	20k	Chest	$(1.5, .75, .75)$	$[-1000, 1000]$	176M[3]	512	$63.38_{\pm 0.05}$	$44.30_{\pm 0.06}$
	Merlin	Vision-language	MerlinAbdCT	15k	Abdomen[4]	$(3.0, 1.5, 1.5)$	$[-1000, 1000]$	121M	2048	$69.80_{\pm 0.05}$	$49.84_{\pm 0.08}$
	COLIPRI	Vision-language, vision-only[5]	CT-RATE, NLST	94k	Chest	$(2.0, 2.0, 2.0)$	$[-1000, 1000]$	147M	768	$73.98_{\pm 0.02}$	$55.45_{\pm 0.10}$
Supervised	**CT-SSG**	None	-	-	-	$(1.5, .75, .75)$	$[-1000, 200]$	12M	512	$74.58_{\pm 0.36}$	$58.75_{\pm 0.28}$

Table 14: Performance comparison of multi-label abnormality classification between CT-SSG (supervised training) and CT-specific foundation models (linear probing), on RAD-ChestCT. *HU range* denotes the Hounsfield Unit range, *Param.* the number of visual encoder parameters (in millions, M), and *Dim.* the dimensionality of the latent representations.
(1) The 69 cohorts are detailed in Table S5 of the CT-FM paper.
(2) Head and neck, Pelvis and Extremity body parts are also included.
(3) The CT-CLIP visual encoder comprises 26M parameters, with an additional 150M parameters in the visual projection head, which maps flattened feature representations to a 512-dimensional embedding space.
(4) Pelvis body parts are also included.
(5) Report generation and opposite sentence losses are also included.

Component	**ResNet3D**	**ViT3D**	**ViViT**	**Swin3D**	**CT-Net**	**CT-MvG**	**CT-Scroll**	**CT-SSG**
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Learning rate	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
(β_1, β_2)	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.99)$	$(0.9, 0.99)$
Batch size	4	4	4	4	4	4	4	4
Warm-up steps	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000
Maximum steps	200,000	200,000	200,000	200,000	200,000	200,000	200,000	200,000
Frozen parameter(s)	None	None	None	None	None	None	None	None
Weights initialization type	Inflation	Inflation	Inflation	Inflation	Direct	Direct	Direct	Direct
Weights initialization backbone	ResNet-18	ViT-S16	ViT-S16	ViT-S16	ResNet-18	ViT-S16	ResNet-18	ResNet-18
Weights initialization source	ImageNet	ImageNet	ImageNet	ImageNet	ImageNet	ImageNet	ImageNet	ImageNet

Table 15: Summary of training protocols for all methods.

Method	Implementation	Model weights
ResNet3D	https://github.com/PPPrior/i3d-pytorch	https://download.pytorch.org/models/resnet18-f37072fd.pth
ViT3D	https://github.com/lucidrains/vit-pytorch	https://huggingface.co/google/vit-base-patch16-224
ViViT	https://github.com/lucidrains/vit-pytorch	https://huggingface.co/google/vit-base-patch16-224
Swin3D	https://github.com/microsoft/Swin3D	https://huggingface.co/microsoft/swin-small-patch4-window7-224
CT-Net	https://github.com/rachellea/ct-net-models	Not available
CT-MvG	https://github.com/compai-lab/2024-miccai-grail-kiechle	Not available
CT-Scroll	https://github.com/theodpzz/ct-scroll	https://huggingface.co/theodpzz/ct-scroll
CT-FM	https://github.com/project-lighter/CT-FM	https://huggingface.co/project-lighter/ct_fm_feature_extractor
CT-CLIP	https://github.com/ibrahimethemhamamci/CT-CLIP	https://huggingface.co/datasets/ibrahimhamamci/CT-RATE
Merlin	https://github.com/StanfordMIMI/Merlin	https://huggingface.co/stanfordmimi/Merlin
COLIPRI	https://huggingface.co/microsoft/colipri	https://huggingface.co/microsoft/colipri

Table 16: Baseline PyTorch implementations and model weights for initialization or direct comparison for backbone initialization, inflation or direct comparison.

Overlap. slices	Nodes	Train	Inference	F1-Score	AUROC
2	119	24 h	105±0.5 ms	57.12±0.21	83.72±0.23
None	80	18 h	70±0.5 ms	57.18±0.19	83.64±0.21

Table 17: Performance comparison between non-overlapping and overlapping triplets of axial slices. *Overlap. slices* refers to the number of overlapping axial slices between two consecutive nodes. *Train* is the training time to reach convergence (in hours, h) and *Inference* the inference time for a sample (in milliseconds, ms) estimated on a NVIDIA RTX A6000 GPU. The two configurations achieve comparable performance, but the overlapping one increases both the training and inference times.

Model		Metrics			
Backbone	Params	F1-Score	AUROC	mAP	Accuracy
CT-RATE					
ResNet18	12M	57.18±0.19	83.64±0.21	58.24±0.50	81.03±0.38
ResNet34	22M	57.84±0.21	83.72±0.23	58.53±0.49	81.36±0.32
RAD-ChestCT					
ResNet18	12M	52.25±0.88	74.58±0.36	58.75±0.28	69.37±1.74
ResNet34	22M	52.48±0.22	75.60±0.31	58.90±0.26	70.76±0.28

Table 18: Comparison of CT-SSG performance using both ResNet-18 and ResNet-34 encoders. The table reports model parameter counts in millions (*Params*, M) and classification performance when node features are extracted using a backbone of increasing capacity, while keeping all other architectural and training settings unchanged. Results show that CT-SSG can benefit from stronger feature encoders, supporting the backbone-agnostic and scalable design of the proposed framework.

Slices per node	Nodes	F1-Score	AUROC	mAP
1	240	56.64±0.26	83.34±0.29	57.31±0.25
3	80	57.18±0.19	83.64±0.21	58.24±0.50
6	40	55.24±0.30	82.22±0.07	55.54±0.10

Table 19: Comparison of CT-SSG performance for different number of axial slices per node. *Slices per node* is the number of adjacent axial slices per node, and *Nodes* is the number of nodes in our graph construction. CT-SSG yields best performance with 3 slices per node, where initial node features are extracted by a 2D ResNet.