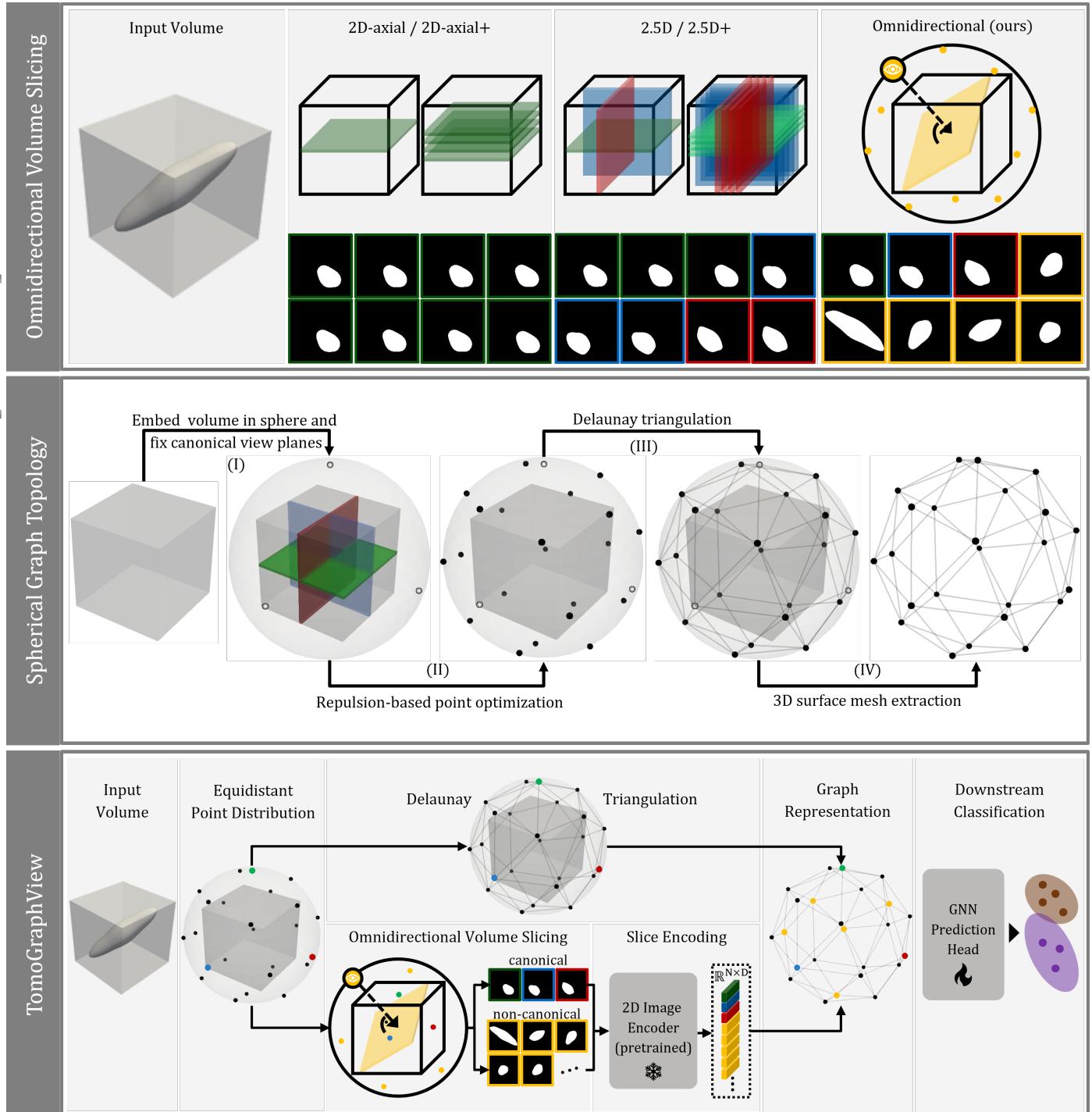


Graphical Abstract

TomoGraphView: 3D Medical Image Classification with Omnidirectional Slice Representations and Graph Neural Networks

Johannes Kiechle, Stefan M. Fischer, Daniel M. Lang, Cosmin I. Bercea, Matthew J. Nyflot, Lina Felsner, Julia A. Schnabel, Jan C. Peeken



Highlights

TomoGraphView: 3D Medical Image Classification with Omnidirectional Slice Representations and Graph Neural Networks

Johannes Kiechle, Stefan M. Fischer, Daniel M. Lang, Cosmin I. Bercea, Matthew J. Nyflot, Lina Felsner, Julia A. Schnabel, Jan C. Peeken

- Novel omnidirectional volume slicing strategy to enable a richer characterization of volumetric structures that are not naturally aligned with canonical planes.
- Novel feature aggregation method based on graph neural networks and spherical graph topology to explicitly model slice relationships.
- Omnidirectional volume slicing outperforms traditional volume slicing methods based on axial, coronal, or sagittal planes across various oncology classification datasets.
- Graph-based feature aggregation outperforms slice-wise feature aggregation baseline methods across various oncology classification datasets.
- Our proposed TomoGraphView framework, integrating omnidirectional slicing and graph-based feature aggregation, surpasses the performance of 3D large-scale pre-trained models.

TomoGraphView: 3D Medical Image Classification with Omnidirectional Slice Representations and Graph Neural Networks

Johannes Kiechle^{a,b,c,g,*}, Stefan M. Fischer^{a,b,c,g}, Daniel M. Lang^{a,c}, Cosmin I. Bercea^{a,c}, Matthew J. Nyflot^f, Lina Felsner^{a,c}, Julia A. Schnabel^{a,c,e,g,**}, Jan C. Peeken^{b,d,**}

^aComputational Imaging and AI in Medicine, School of Computation, Information and Technology, Technical University of Munich, Germany
^bDepartment of Radiation Oncology, TUM School of Medicine, TUM University Hospital rechts der Isar, Technical University of Munich, Germany

^cInstitute of Machine Learning in Biomedical Imaging, Helmholtz Center Munich, Neuherberg, Germany

^dInstitute of Radiation Medicine, Helmholtz Center Munich, Neuherberg, Germany

^eComputational Imaging, School of Biomedical Engineering and Imaging Sciences, King's College London, United Kingdom

^fDepartment of Radiation Oncology, University of Washington, Seattle, United States

^gMunich Center for Machine Learning, Germany

Abstract

The growing number of medical tomography examinations has necessitated the development of automated methods capable of extracting comprehensive imaging features to facilitate downstream tasks such as tumor characterization, while assisting physicians in managing their growing workload. However, 3D medical image classification remains a challenging task due to the complex spatial relationships and long-range dependencies inherent in volumetric data. Training models from scratch suffers from low data regimes, and the absence of 3D large-scale multimodal datasets has limited the development of 3D medical imaging foundation models. Recent studies, however, have highlighted the potential of 2D vision foundation models, originally trained on natural images, as powerful feature extractors for medical image analysis. Despite these advances, existing approaches that apply 2D models to 3D volumes via slice-based decomposition remain suboptimal. Conventional volume slicing strategies, which rely on canonical planes such as axial, sagittal, or coronal, may inadequately capture the spatial extent of target structures when these are misaligned with standardized viewing planes. Furthermore, existing slice-wise aggregation strategies rarely account for preserving the volumetric structure, resulting in a loss of spatial coherence across slices. To overcome these limitations, we propose *TomoGraphView*, a novel framework that integrates *omnidirectional* volume slicing with *spherical graph-based feature aggregation*. Unlike traditional methods, which are restricted to canonical views, our approach samples both canonical and non-canonical cross-sections. These non-canonical views are derived from uniformly distributed points on a sphere, which visually encompasses the 3D volume, thereby producing a richer set of cross-sectional representations. As the spherically distributed viewpoints naturally define a spherical graph topology after triangulation, we allow for the explicit encoding of spatial relationships across views as nodes and corresponding edges in the underlying graph topology, and leverage a graph neural network for spatial-aware feature aggregation. Experiments across six oncology 3D medical image classification datasets demonstrate that *omnidirectional* volume slicing improves the average performance in Area Under the Receiver Operating Characteristic Curve (AUROC) from 0.7701 to 0.8154 compared with traditional slicing approaches relying on canonical view planes. Moreover, we can further improve AUROC performance from 0.8198 to 0.8372 by leveraging our proposed graph neural network-based feature aggregation. Notably, *TomoGraphView* surpasses large-scale pretrained 3D medical imaging models across all datasets and tasks, underscoring its effectiveness as a powerful framework for volumetric analysis and therefore represents a key step toward bridging the gap until fully native 3D foundation models become available in medical image analysis.

Keywords: 3D Medical Image Classification, Omnidirectional Volume Slicing, Graph Neural Networks

1. Introduction

Deep learning methods have emerged as promising tools in medical imaging research for their ability to learn complex, hierarchical feature representations from large imaging datasets. Such approaches enable the detection of pathological processes, supporting applications ranging from early disease detection and accurate diagnosis [56], malignancy prediction [54, 47, 48],

and lesion characterization [29, 86, 16], to robust image segmentation [58, 33, 10, 21] and prognostic as well as predictive modeling [89, 62, 65, 39]. These capabilities not only facilitate personalized treatment strategies but also hold the potential to improve therapeutic effectiveness and patient outcomes [57]. Driven by technological advancements that have led to a wealth of tomography data, image analysis methods have gained increasing relevance in medical research, demanding automated tools to support clinicians [70]. To this end, the automated and effective extraction of imaging biomarkers from 3D medical imaging data constitutes a core component that substantially al-

*Corresponding author: johannes.kiechle@tum.de

**These senior authors contributed equally to this work.

leviates clinical workload, while also enhancing the efficacy of subsequent downstream tasks [57].

Generally, a promising direction to extract semantically rich imaging features from 2D images is represented by foundation models, which have already attracted significant attention in deep learning based computer vision applications [55, 37, 90, 50, 22, 72, 75, 63]. They not only show great promise in their ability to extract comprehensive feature representations from imaging data but also in their easy adaptability to a wide range of downstream tasks, such as segmentation [4], classification [73], or registration [64]. While applying these powerful models has become standard practice in natural image analysis, the progress in adapting them to 3D medical imaging is proceeding more slowly. A key challenge is the lack of large and publicly available multimodal medical imaging datasets required for training at scale [5]. Following Paschali et al. [52], a foundation model in radiology should incorporate both large-scale model architectures and multimodal training data, leverage self-supervised training to reduce the need for expert annotations, and crucially, demonstrate emergent capabilities beyond their training objectives. Despite the progress and rapid development of large-scale pretrained 3D models in medical imaging research, performance levels of radiology foundation models still lag behind their 2D counterparts [74]. Regardless of challenges such as the limited availability of large-scale 3D datasets and the computational demands of training 3D models, ongoing research in these areas is likely to lead to substantial breakthroughs and enhanced integration of these models into clinical workflows [5, 74].

Despite 3D large-scale pretrained models lagging behind the performance levels of their 2D counterparts, recent studies have demonstrated that 2D vision foundation models trained on natural images can be effectively transferred to medical imaging tasks [32, 42]. In particular, the DINO [11] family of models has shown considerable promise in the medical imaging domain [31, 15, 6, 53]. However, applying 2D models to 3D input data necessitates slice-based decomposition of the input volumes, resulting in the loss of structural information across slices, as slice-wise representations disregard the three-dimensional spatial context inherent in 3D volumes. A common strategy to combine the resulting slice-wise information extracted by a pretrained encoder is to use a dimension-wise average thereof or a shallow multi-layer perceptron (MLP) [36, 46, 42]. While this enables end-to-end learning for downstream tasks such as classification, there is no inherent mechanism that explicitly accounts for the relative positions among the slices. To address this limitation, more structured approaches have been explored. Long short-term memory (LSTM)-based aggregation methods [20, 49, 41] treat consecutive slices as temporal sequences, thereby capturing inter-slice relationships. Alternatively, transformer-based aggregation [46, 44] leverages self-attention mechanisms to model spatial relationships across slices more directly.

With the emergence of graph neural networks (GNNs) [61], effective modeling of naturally graph-structured data has become possible. GNNs are particularly well-suited for problems involving pairwise interactions or modeling complex relational

data by encoding relationships in the graph structure [82]. In the context of 3D data, GNNs have demonstrated a strong ability to preserve an underlying spatial structure when the graph topology is based on the relative position of 2D representations [79, 80]. This capability can be extended to 3D medical data, where GNNs can effectively represent spatial relationships between encoded slices within the input volume [36, 18].

In this work, we present *TomoGraphView*, a novel framework designed to preserve the spatial structure of volumetric medical imaging data when decomposed into 2D slices compatible with 2D pretrained encoders. The name reflects its core components: *Tomo* refers to tomography data, *Graph* denotes the use of GNN-based feature aggregation, and *View* signifies the integration of multi-view information from *omnidirectional* perspectives. *TomoGraphView* consists of two main components: *Omnidirectional volume slicing* and subsequent *spherical graph-based feature aggregation*. In our approach, volumetric data, such as CT or MRI scans, are represented as structured graphs, with each graph corresponding to a single tomography volume. The nodes represent features encoded by a 2D vision foundation model given cross-sectional views, covering both canonical and non-canonical planes. The graph topology is defined by triangulating these viewpoints, thereby approximating a 3D spherical structure. This design enables a more faithful representation of the 3D volume by embedding both the 2D slice information and their spatial relationships within the three-dimensional graph. Edges between nodes are weighted according to their relative distance, allowing the model to efficiently integrate local and global context while explicitly preserving spatial structure. By aggregating node features through a GNN, our framework effectively models 3D context from 2D representations, demonstrating the strength of graph-based feature aggregation for volumetric data analysis. The main contributions and findings of this work can be summarized as follows:

(I) We introduce a novel out-of-plane volume slicing strategy, termed *omnidirectional volume slicing*, which generalizes beyond conventional axial, coronal, and sagittal slicing and demonstrates superior downstream performance when compared to traditional volume slicing approaches solely relying on canonical axes.

(II) We introduce a spherical mesh-based graph structure derived from our omnidirectional slicing strategy, enabling the preservation of spatial relationships between extracted 2D slices from tomographic data through spherical graph-based feature aggregation.

(III) Extensive experiments on six oncology 3D medical imaging datasets demonstrate that the proposed *TomoGraphView* framework, integrating *omnidirectional* volume slicing with *spherical graph-based feature aggregation*, consistently outperforms baseline slice-wise aggregation methods and large-scale pretrained 3D medical models across all datasets and tasks, establishing *TomoGraphView* as a powerful framework for 3D medical image classification.

2. Related Works

In the context of preserving the spatial structure of volumetric data when decomposed into 2D slices compatible with 2D encoders, we review prior work in two main areas. In Section 2.1, we begin with 2D slice-based methods for 3D medical image analysis, focusing specifically on approaches that extract slices directly from volumetric data. Projection-based methods, such as maximum intensity projections [3, 30, 67], which collapse 3D volumes into 2D representations, are therefore excluded from this discussion. Subsequently, in Section 2.2, we turn our attention to slice-wise feature aggregation methods.

2.1. Slice-based Methods in 3D Medical Imaging

Two-dimensional slice-based models continue to be an attractive alternative to heavy 3D models, as they are computationally lighter, benefit from the decomposition of 3D volumes into numerous 2D slices, and can leverage powerful pretrained backbones from natural image analysis, which have already demonstrated strong performance levels in medical image classification tasks [32, 42].

Several studies have demonstrated the effectiveness of 2D approaches. Navarro et al. [48] fine-tuned a DenseNet161 pre-trained on ImageNet using axial T1- and T2-weighted MRI slices of soft-tissue sarcomas for tumor grading. Similarly, Ahamed et al. [2] trained a ResNet18 on axial PET/CT slices of lymphoma to classify whether a slice intersected tumor tissue. Wang et al. [77] proposed a two-stage slice-to-volume feature representation framework for Alzheimer’s disease, where informative axial MRI slices were selected and aggregated into volume-level features. In parallel, self-supervised foundation models have also been evaluated for slice-level tasks. Huang et al. [31] and Baharoon et al. [6] applied DINOV2 to axial brain MRIs and chest CTs for glioma grading and COVID-19 diagnosis, respectively, while Liu et al. [42] assessed DINOV3 on non-contrast axial CT slices for detecting 18 clinically significant abnormalities.

Beyond single-plane analysis, multi-plane strategies that combine axial, coronal, and sagittal slices, commonly referred to as 2.5D, have been proposed. Saint-Estèven et al. [60] employed an Xception model trained on 2.5D representations to predict HPV status in oropharyngeal cancer. Similarly, Meneghetti et al. [44] explored foundation model-based multiple instance learning for head and neck cancer outcomes, comparing single-plane, multi-plane, and 3D sub-volume representations. A related hybrid approach by Jang and Hwang [34] first extracted volumetric features using a 3D CNN, followed by multi-plane slicing of the feature tensor and 2D CNN and transformer processing for Alzheimer’s disease classification.

However, none of these methods extract slices from non-canonical orientations, i.e., directions that are not aligned with the standard axes. Consequently, valuable information from structures misaligned with these planes may be omitted. Moreover, while choosing a powerful method for decomposing 3D data into 2D slices is essential, the aggregation of slice-wise

features remains equally critical for preserving volumetric context and achieving strong downstream performance, as discussed in the following section.

2.2. Slice-wise Feature Aggregation in 3D Medical Imaging

Slice-wise feature aggregation represents a promising direction for 3D medical image analysis. By leveraging powerful 2D foundation models as feature encoders, it avoids the high computational cost of native 3D training while allowing a flexible combination of slice embeddings. However, decomposing a 3D volume into 2D slices inherently discards volumetric structure, meaning that spatial dependencies between slices must be explicitly modeled to preserve the structural integrity of the original data.

The simplest approach aggregates slice embeddings into a single volume-level representation using non-learnable operations such as mean pooling. Müller-Franzes et al. [46] employed this strategy by encoding axial slices with DINOV2 and averaging embeddings to classify breast MRI, lung CT, and knee MRI scans across multiple tasks. Similarly, Liu et al. [42] extracted DINOV3-based embeddings from axial chest CT slices to classify anatomical structures into 18 clinical categories. Although efficient, such mean-pooling schemes ignore inter-slice relationships and therefore likely yield suboptimal performance.

To introduce learnability, several studies replaced pooling with multi-layer perceptron (MLP) aggregation. Truong et al. [71] fine-tuned multiple self-supervised ImageNet models with an MLP head for various medical classification tasks involving lymph node, fundus, and chest X-ray images. Baharoon et al. [6] extended this idea to 100 radiology experiments across modalities (X-ray, CT, MRI), while Kiechle et al. [36] compared MLP-based aggregation across anatomical structures in the MedMNIST3D dataset. Although MLPs provide learnable fusion, they typically neglect slice order and spatial relations unless explicitly conditioned on slice indices [36].

Recurrent architectures such as long short-term memory (LSTM) networks address this limitation by treating slices as ordered sequences, thus maintaining spatial continuity [20]. Zhang et al. [87] introduced a slice-wise attention network that models MRIs as sequential inputs to capture long-range dependencies for multiple sclerosis lesion segmentation. Similarly, Nguyen et al. [49] combined CNNs with LSTMs for intracranial hemorrhage detection in CT scans, while Lilhore et al. [41] integrated bidirectional LSTMs with EfficientNet-B0 for breast cancer classification. Miron et al. [45] also processed volumes slice-by-slice through a CNN, fused feature maps at the channel level, and refined them using a lightweight convolutional head into a prediction.

More recently, transformer-based models have gained traction by leveraging attention to learn inter-slice dependencies. Jang and Hwang [34] proposed a multi-plane and multi-slice transformer (M3T) for Alzheimer’s diagnosis, fusing axial, coronal, and sagittal features extracted from a 3D CNN. Meneghetti et al. [44] introduced an end-to-end multiple instance learning (MIL) transformer for predicting clinical outcomes in head and neck cancer, encoding 2D CT slices with

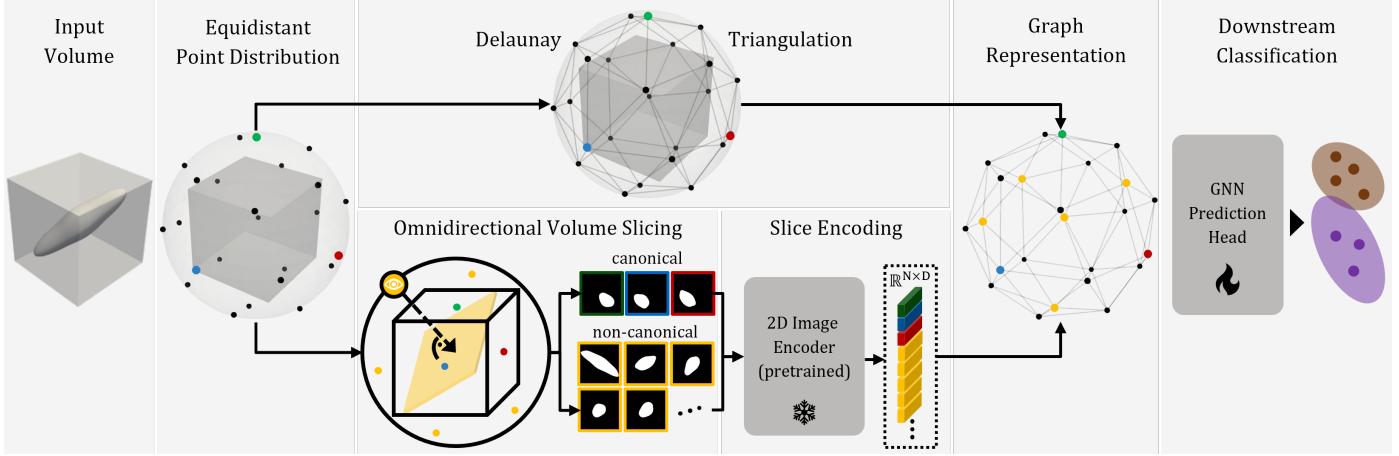


Figure 1: Visual overview of the proposed *TomoGraphView* framework. The input volume is enclosed within a sphere with N equidistant surface points, three corresponding to canonical planes indicated as green (axial), blue (coronal), red (sagittal), and the remaining $N - 3$ distributed via repulsion-based optimization. From this distribution, two processing branches emerge: Delaunay triangulation forms a spherical mesh defining the graph topology (done only once, for any N node graph topology), and omnidirectional volume slicing generates canonical and non-canonical slices along the directions of the sampled points. The resulting slices are encoded into 1D vector representations that serve as node features within the spherical graph. A GNN prediction head performs the downstream classification.

a foundation model and aggregating them using transformer-based attention. Müller-Franzes et al. [46] later presented the Medical Slice Transformer (MST), which combines transformer aggregation with 2D self-supervised encoders and demonstrates strong results across breast MRI, lung CT, and knee MRI classification.

A complementary direction represents volumetric data as structured graphs. Kiechle et al. [36] and Di Piazza et al. [18] modeled tomographic volumes where nodes correspond to slice features and edges encode spatial adjacency. Kiechle et al. [36] showed that graph representations improve accuracy and runtime over MLP-based aggregation but depend heavily on selecting suitable graph convolution operators and topologies, limiting general applicability. Di Piazza et al. [18] constructed graphs from axial slice stacks to capture both local and global inter-slice dependencies along the axial axis. However, this design remains restricted to a single orientation, omitting valuable coronal and sagittal context. Moreover, evaluation limited to chest CTs leaves the generalizability of this approach across modalities and anatomical sites uncertain.

3. Methodology

In this work, we present *TomoGraphView*, a novel framework that preserves the spatial structure of volumetric medical imaging data when decomposed into 2D slices compatible with 2D feature encoders. The name *TomoGraphView* encapsulates its core components: *Tomo* refers to tomography data, *Graph* highlights the GNN-based feature aggregation, and *View* emphasizes the integration of multi-view information from *omnidirectional* perspectives. In general, our framework is built upon two key components: *omnidirectional volume slicing* and *spherical mesh-based graph topology* construction with corresponding GNN-based feature aggregation. A visual summary of the method is depicted in Figure 1. The following sections detail

each component of *TomoGraphView*. We first describe omnidirectional volume slicing in Section 3.1 and relate it to conventional slicing approaches. Section 3.2 outlines the encoding of the obtained omnidirectional 2D cross-sectional slices into 1D feature representations. In Section 3.3, we introduce the *spherical mesh-based graph topology* construction, followed by Section 3.4, which describes the graph representation learning process used for downstream classification.

3.1. Omnidirectional Volume Slicing

Traditional volume slicing strategies typically rely on canonical planes such as axial, coronal, and sagittal views, or a combination thereof, referred to as 2.5D. However, before decomposing a volume into 2D slices, it is common practice in 3D medical image classification to first extract a subvolume enclosing the target structure of interest, commonly defined using a segmentation mask. Accordingly, the most representative slice along each direction is defined as the one containing the largest visible lesion area. As illustrated in Figure 2, in the 2D-axial setting, a single slice is extracted along the axial direction. The 2D-axial+ variant further includes adjacent slices symmetrically around the axial largest-lesion slice, resulting in a total of N views. The 2.5D approach combines information from all the canonical planes by selecting the most representative slice along the axial, coronal, and sagittal directions, whereas 2.5D+ extends this strategy by additionally including adjacent slices symmetrically around the largest-lesion slice in each direction.

In addition to utilizing information from canonical planes, we propose to extract slices from non-canonical orientations to obtain *omnidirectional* slice representations, thereby enabling a richer characterization of volumetric structures that are not naturally aligned with the canonical axes. To obtain N *omnidirectional* slice representations, we visually embed the volume within a bounding sphere \mathcal{S} , centered at the origin, which is formally described as

$$\mathcal{S} = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\|_2 \leq r\}, \quad (1)$$

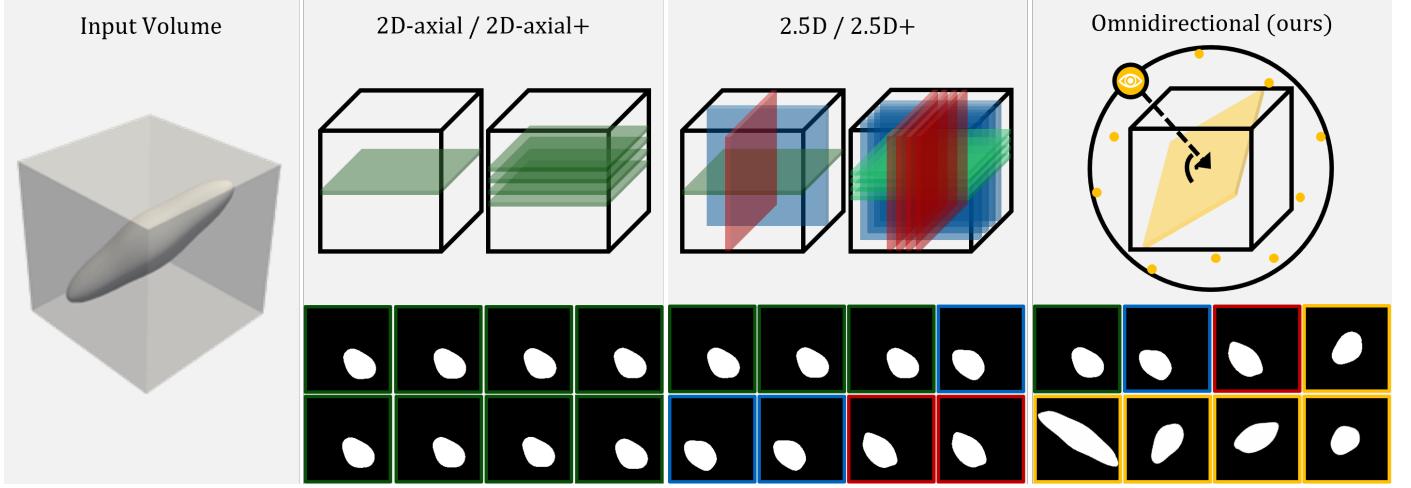


Figure 2: Visual representation of different volume slicing strategies and resulting slices for a synthetically generated input volume depicting an elongated shape diagonally within the volume. Volume slicing strategies include 2D-axial and 2D-axial+ (left), 2.5D and 2.5D+ (middle), and our proposed omnidirectional volume slicing (right). Green slices represent slices extracted from the axial direction, red from the sagittal direction, and blue from the coronal direction, whereas yellow shows out-of-plane slices as employed within our *omnidirectional* volume slicing strategy.

where r is the radius such that the sphere fully contains the 3D volumetric scan as illustrated in Figure 3 (step I). Next, as we intend to make use of both, canonical and non-canonical perspectives, we first of all fix three points on the surface of the sphere representing the canonical view planes (i.e., axial, coronal, and sagittal) with their known coordinates $(r, 0, 0)$, $(0, r, 0)$ and $(0, 0, r)$, as illustrated in Figure 3 (step I).

We then optimize the location of the remaining $(N-3)$ points, representing non-canonical view planes, such that all points on the surface are approximately equally distributed, as illustrated in Figure 3 (step II). To this end, we make use of the repulsion-based optimization outlined in Algorithm 1, which is inspired by the Thomson problem [69]. Therein, points are interpreted as identical charged particles constrained to lie on the spherical surface, and their positions are optimized by minimizing the system’s total electrostatic energy. Formally, given N points $\{x_i\}_{i=1}^N$ with $x_i \in \mathbb{R}^3$ and $\|x_i\| = 1$, the pairwise Coulomb energy is defined as

$$E = \sum_{i \neq j} \frac{1}{\|x_i - x_j\|^2}. \quad (2)$$

The gradient of this energy function yields repulsive forces between all point pairs which is used to iteratively update the position of the $N - 3$ non-fixed points. After each update step, points are re-projected onto the sphere to enforce the unit-norm constraint. By visually enclosing the volume in a sphere, both centered at the origin, we guarantee that all possible slice orientations, which are defined by points on the sphere, will intersect the volume. For each volume, canonical and non-canonical slices are then extracted by first aligning the slicing plane such that its normal vector points to the respective point sampled on the surface of the sphere, and second by identifying the largest lesion slice from the respective canonical and non-canonical plane. Repeating this process for all $i = 1, \dots, N$ yields a set of N uniformly distributed omnidirectional slices, thereby effectively resembling the spatial structure of the input volume.

Algorithm 1: Repulsion-based point optimization

```

Input:  $N$  total points,  $F$  fixed points on sphere
Place fixed points at given coordinates;
Randomly sample  $(N - F)$  additional points on sphere;
while not converged do
  foreach point  $i$  do
    if  $i$  is fixed then
      continue;
    Compute Coulomb force from all other points
     $j \neq i$ :
     $F_i \leftarrow \sum_{j \neq i} \frac{1}{\|x_i - x_j\|^2}$ ;
    Update position:  $x_i \leftarrow x_i + \eta \cdot F_i$ ;
    Reproject:  $x_i \leftarrow x_i / \|x_i\|$ ;

```

Output: Optimized set of N points on the sphere

3.2. Cross-sectional Image Encoding

After omnidirectional volume slicing, we transfer the obtained 2D volume slices into 1D feature representations by means of DINOv2 [50]. Generally, the DINO [11] framework is a self-supervised learning method based on a student–teacher architecture designed to extract semantically rich and meaningful visual representations. Building on this, Oquab et al. [50] introduced DINOv2, an extension of DINO [11] trained with self-supervised contrastive learning on a large-scale dataset of 142 million curated images. This large-scale training enables DINOv2 to learn highly transferable semantic features, which have demonstrated strong performance across diverse vision tasks, including both pixel-level tasks (e.g., segmentation and detection) and image-level tasks such as classification. In our framework, we employ a pretrained DINOv2 vision transformer (ViT) model in its small variant, featuring 21M parameters. The encoder is kept frozen during all experiments and processes input slices of size 224×224 pixels, obtained by lin-

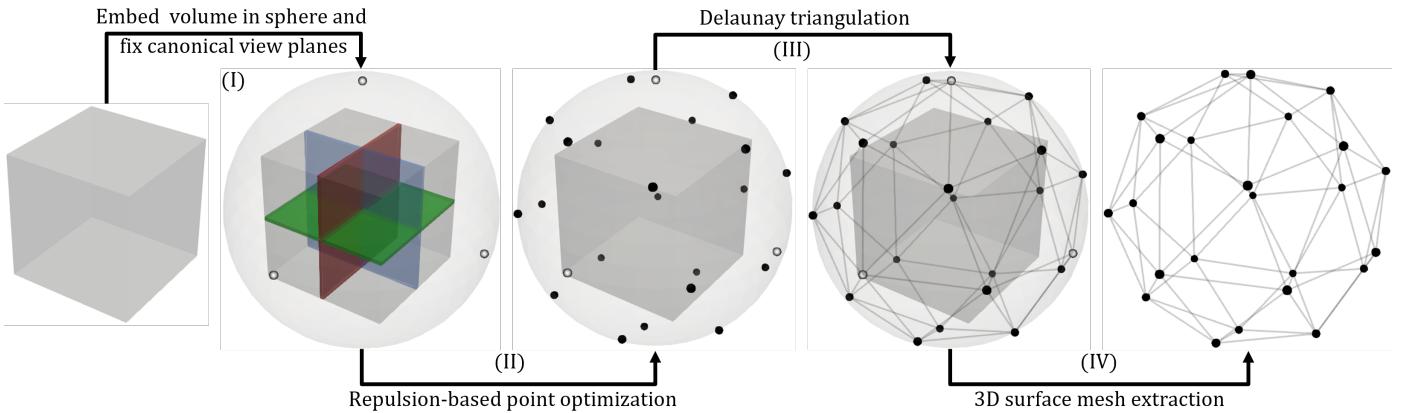


Figure 3: Sequential overview of our spherical mesh-based graph topology construction: (I) We fix canonical view planes with coordinates $(1, 0, 0), (0, 1, 0)$ and $(0, 0, 1)$ on the surface of the sphere, (II) we add additional $(N - 3)$ points on the sphere and optimize their position for equidistant distribution across the surface, (III) we apply Delaunay triangulation to establish edge connection between neighboring points, and (IV) extract the 3D surface mesh.

early up-sampling the input images if not matched initially. The model produces a 384-dimensional feature representation for a single input image corresponding to a single view.

3.3. Spherical Mesh-based Graph Topology Construction

Given the N view points, which are equidistantly distributed on the spherical surface (see Section 3.1) and their corresponding cross-sectional volume slice encodings (see Section 3.2), we proceed by describing the construction of the *spherical mesh-based graph topology*. Therein, nodes represent canonical and non-canonical view planes and slice encodings their corresponding feature vector. We denote the set of points uniformly distributed on the spherical surface as $\mathbf{P} = \{\mathbf{p}_1, \dots, \mathbf{p}_N\}$. To derive a graph structure from the sampled points, we apply Delaunay triangulation to \mathbf{P} , which guarantees that no point lies inside the circumcircle of any spherical triangle [17]. The triangulation defines the edge set

$$\mathbf{E} = \{(i, j) \mid \mathbf{p}_i, \mathbf{p}_j \text{ are adjacent in the triangulation}\}, \quad (3)$$

yielding an undirected graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{A})$, where the vertices $\mathbf{V} = \mathbf{P}$ and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the weighted adjacency matrix, where all connections \mathbf{e}_{ij} resulting from the triangulation have an edge weighting $\mathbf{A}_{ij} = \mathbf{w}_{ij} = 1$. This procedure results in a mesh-based graph topology with near-uniform node distribution and well-defined local neighborhoods. Two properties that have been proven to be advantageous for subsequent mesh-based graph learning tasks [8]. A graphical visualization of the Delaunay triangulation (step III) and 3D surface mesh extraction (step IV) can be found in Figure 3. Besides the local connectivity pattern, where points \mathbf{p}_i and \mathbf{p}_j are connected when they are adjacent in the triangulation, we also introduce weighted cross-connections between all node pairs, effectively forming a complete graph. The corresponding edge weights \mathbf{w}_{ij} of cross-connections are defined as the inverse node distance. Here, the distance refers to the hop distance (i.e., shortest path length between two nodes $(\mathbf{v}_i, \mathbf{v}_j \in \mathbf{V})$ using Dijkstra algorithm [19], defined as

$$\mathbf{d}_{\mathbf{G}}(\mathbf{v}_i, \mathbf{v}_j) = \min_{\mathbf{p} \in \mathbf{P}(\mathbf{v}_i, \mathbf{v}_j)} |\mathbf{p}|, \quad (4)$$

where $\mathbf{p} \in \mathbf{P}(\mathbf{v}_i, \mathbf{v}_j)$ is the set of all paths from node \mathbf{v}_i to node \mathbf{v}_j , and $|\mathbf{p}|$ refers to the number of edges in path \mathbf{p} , which shows the minimum distance from node \mathbf{v}_i to node \mathbf{v}_j . Thus, $\mathbf{d}_{\mathbf{G}}(\mathbf{v}_i, \mathbf{v}_j)$ corresponds to the minimum number of hops required to reach node \mathbf{v}_j from node \mathbf{v}_i . This finally results in edge weights computed as

$$\mathbf{w}_{ij} = \frac{1}{\mathbf{d}_{\mathbf{G}}(\mathbf{v}_i, \mathbf{v}_j)} = \frac{1}{(\min_{\mathbf{p} \in \mathbf{P}(\mathbf{v}_i, \mathbf{v}_j)} |\mathbf{p}|)} \quad (5)$$

3.4. Graph Representation Learning

Let $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{A})$ denote a spherical mesh-based graph, as introduced in Section 3.3, with node attributes \mathbf{X}_v for $v \in \mathbf{V}$ and edge weights $\mathbf{w}_{ij} = \mathbf{A}_{ij}$ for $(i, j) \in \mathbf{E}$. Given a set of graphs $\{\mathbf{G}_N\}$ and their labels $\{\mathbf{y}_N\}$, the task of graph supervised learning is to learn a representation vector $\mathbf{h}_{\mathbf{G}} \in \mathbb{R}^n$ that is used to predict the class of the entire graph. Generally, graph representation learning can be derived from three different flavors of GNN layers, namely convolutional, attentional, and message-passing [7]. In this work, we focus on GNNs that follow message-passing, as it provides a flexible and conceptually intuitive framework where node features are aggregated among neighborhoods, following the underlying graph structure [24]. For a single iteration, this is generically represented as follows

$$\mathbf{h}_i = \phi \left(\mathbf{x}_i, \bigoplus_{j \in \mathcal{N}_i} \psi(\mathbf{x}_i, \mathbf{x}_j) \right), \quad (6)$$

where \mathbf{x}_i denotes the feature vector of node \mathbf{v}_i , and \mathcal{N}_i represents its set of neighboring nodes. The function $\psi(\mathbf{x}_i, \mathbf{x}_j)$ computes a message from neighbor j to node i , and the operator \bigoplus aggregates all incoming messages using a permutation-invariant function. The update function ϕ combines the i -th node feature \mathbf{x}_i with the aggregated neighborhood representation to produce the updated embedding \mathbf{h}_i . More specific for GraphSAGE [26],

the updated node representation \mathbf{h}_i can be described as

$$\mathbf{h}_i = \phi \left(\mathbf{W} \left[\mathbf{x}_i \| \bigoplus_{j \in N_i} (\{\mathbf{e}_{ji} \cdot \mathbf{x}_j, \forall j \in N(i)\}) \right] \right), \quad (7)$$

where \mathbf{W} is a learnable weight matrix, $\|$ represents the concatenation operation, and \mathbf{e}_{ji} is the scalar weight on the edge from node j to node i . As using multiple node neighborhood aggregation functions has been shown to provide notable improvements [14], we leverage both mean and max aggregation. They are particularly suitable to capture the distribution of elements within the graph (i.e., mean) and prove to be advantageous to identify representative elements (i.e., max) [83]. As our underlying graph topology does not change within a setting of N nodes, we do not use sum aggregation, which has been shown to be especially powerful for learning structural graph properties. After the node updates, the resulting set of N one-dimensional node feature vectors is aggregated using dimension-wise mean pooling δ (i.e., a mean readout function) to obtain a global graph representation $\mathbf{h}_G \in \mathbb{R}^{384}$. This representation is then passed through a linear layer σ to predict the binary class label \mathbf{y}_G for the entire graph, formally expressed as

$$\mathbf{y}_G = \sigma(\mathbf{h}_G), \quad \text{where } \mathbf{h}_G = \delta(\{\mathbf{h}_v \mid v \in G\}). \quad (8)$$

4. Experiments & Results

We begin our experimental evaluation section by first describing the datasets used in this study (Section 4.1), followed by implementation and training details (Section 4.2). We then benchmark our proposed volume slicing strategy in Section 4.3 before performing an evaluation on our *TomoGraphView* framework. We further extend our experimental analysis by investigating the impact of volume slice spacing (Section 4.5), a graph topology ablation study (Section 4.6), and other slice-wise feature aggregation methods (Section 4.7). Finally, we benchmark our proposed *TomoGraphView* framework against 3D pretrained models (Section 4.8), comparing 3D frozen backbone and 3D model fine-tuning performances.

4.1. Datasets

In our experimental evaluation, we concentrate on six clinically relevant oncological tasks. A comprehensive overview of the datasets used is provided below, while Figure 4 presents a visual summary of these datasets.

Brain Tumors [9] This open-source dataset comprises $n = 501$ patients with brain gliomas, with a median voxel spacing of $1.00 \times 1.00 \times 1.00$ mm. It contains T1-weighted post-contrast (T1c) MRI scans from patients diagnosed with glioblastoma, oligodendrogloma, or astrocytoma. The cohort was collected at a single institution, the University of California, San Francisco, USA. For our analysis, tumors were categorized into high-grade (G4) and low-grade (G2/G3) according to the WHO CNS classification.

Head-Neck Tumors [78, 25, 40] This open-source dataset comprises CT scans of $n = 545$ patients with head and neck

squamous cell carcinoma (HNSCC) of the oropharynx, acquired at a median voxel spacing of $0.98 \times 0.98 \times 2.00$ mm. The multi-institutional cohort was assembled from Maastricht University Medical Center, Netherlands, University of Texas MD Anderson Cancer Center, USA, and Princess Margaret Cancer Center, Ontario, Canada. The task involves binary classification of human papillomavirus (HPV) status into HPV-positive and HPV-negative entities.

Breast Tumors [59] This open-source dataset includes dynamic contrast-enhanced (DCE) MRI scans of $n = 192$ patients with biopsy-confirmed invasive breast cancer, acquired at a median voxel spacing of $0.74 \times 0.74 \times 1.00$ mm. The single-institutional cohort was retrospectively collected at Duke Hospital, North Carolina, USA. Tumors were stratified into high-grade (“high”) and low-grade (“intermediate”/“low”) categories according to their Nottingham grade. In our experiments, we use the first post-contrast DCE-MRI volumes.

Liver Tumors [43] This open-source dataset consists of contrast-enhanced multi-phase CT scans from $n = 176$ patients with primary liver tumors, acquired at a median voxel spacing of $1.00 \times 1.00 \times 1.00$ mm. The cohort was collected at the Radiology Department of Chongqing Yubei District People’s Hospital, China. Binary classification was performed on the venous phase (C2) to distinguish hepatocellular carcinoma (HCC) from combined hepatocellular–cholangiocarcinoma (cHCC-CCA), the latter being clinically more aggressive due to poorer overall survival at comparable stage.

Kidney Tumors [28] This open-source dataset contains contrast-enhanced multi-phase CT scans of $n = 210$ patients with kidney tumors, acquired at a median voxel spacing of $0.78 \times 0.78 \times 3.0$ mm. The single-institutional cohort was retrospectively collected at the University of Minnesota, USA. Tumors were graded into high-grade (G3/G4) and low-grade (G1/G2) entities according to the WHO ISUP classification, as determined from post-operative surgical pathology reports.

Soft-Tissue Tumors This in-house dataset comprises MRI scans of $n = 291$ patients with soft-tissue sarcomas (STS) of the extremities and trunk, acquired with T2-weighted fat-saturation (T2FS) sequences at a median voxel spacing of $0.78 \times 0.94 \times 4.80$ mm. Two independent patient cohorts were retrospectively collected at the University of Washington/Seattle Cancer Care Alliance, USA, and the Technical University of Munich, Germany. Tumors were categorized into high-grade (G2/G3) and low-grade (G1) entities based on histopathological grading.

4.2. Implementation and Training Details

All volumes were first aligned to a common positive Right–Anterior–Superior (RAS+) orientation and subsequently resampled to an isotropic voxel spacing of $1 \times 1 \times 1$ mm. In order to minimize distortion when resizing slices to match the DINov2 input resolution, we applied symmetric cropping around the lesion. For our proposed *omnidirectional* slicing strategy, scalar image volumes were resampled using linear interpolation, while segmentation masks were resampled with nearest-neighbor interpolation to preserve their categorical labels. For

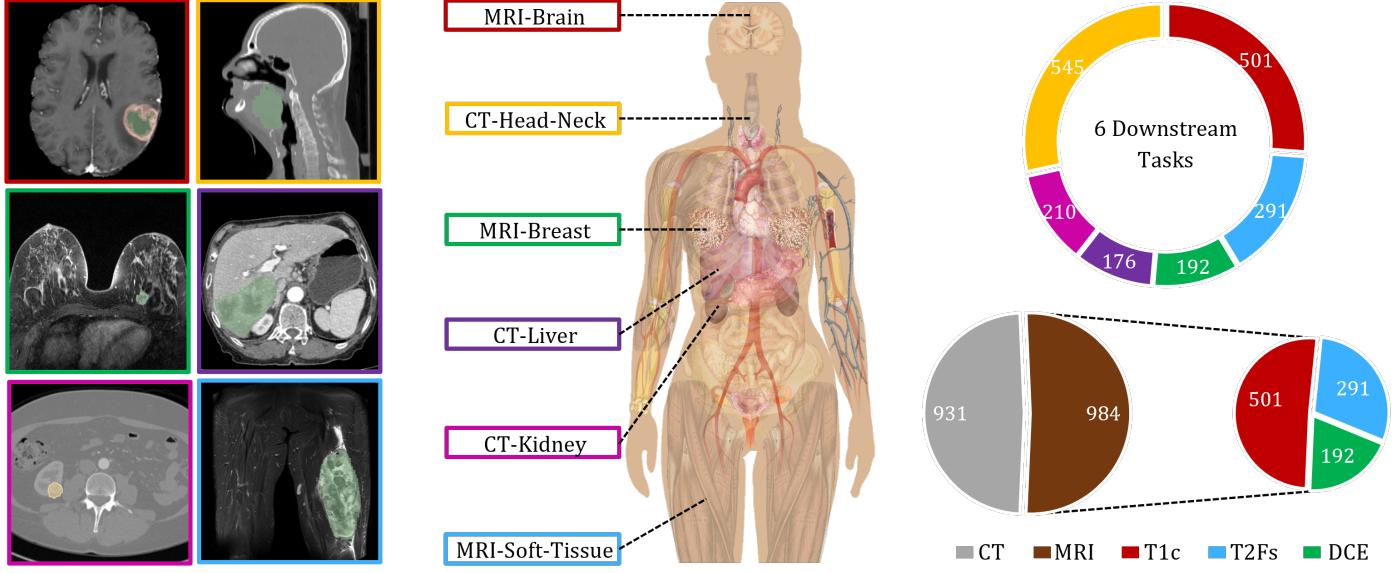


Figure 4: Visual summary of the collected datasets from various target structures among the human body, including a visual example from each dataset and the number of images under different acquisition criteria. Here, CT refers to computed tomography, MRI refers to magnetic resonance imaging, T1c refers to T1-weighted post-contrast MRI, T2Fs refers to T2-weighted fat-saturated MRI, and DCE refers to dynamic contrast-enhanced MRI.

all experiments, we attach a lightweight classification head to the frozen DINOv2 feature extractor. Specifically, the head of each comparison method consists of two layers with a nonlinear ReLU activation in between, featuring a total of 100k trainable parameters to ensure a fair comparison across methods. We restrict the prediction head to a shallow design to avoid overfitting, given the limited dataset sizes in medical imaging, which also helps to mitigate graph oversmoothing in the case of GNNs [35]. For the MLP, slice encodings are concatenated. We train and evaluate our models using a stratified 5-fold cross-validation scheme according to the binary class label. More specifically, we use three splits for training, one for validation, and one for testing. Training is conducted using the SGD optimizer with a weight decay of $1e^{-3}$, a batch size of 16, and 300 epochs. The model that achieves the best AUROC score on the validation set is selected for final evaluation on the test set. A learning rate scheduler first linearly increases the learning rate to 0.001 for 100 warm-up epochs, and decays the learning rate afterwards by a factor of 0.95 when there is no improvement in AUROC validation score for at least 5 epochs, to ensure smooth convergence. All models were trained on an Nvidia RTX A6000. We publicly share our code base¹ and provide a user-friendly library for omnidirectional volume slicing².

4.3. Volume Slicing Strategy Benchmarking

In this experiment, we compare the baseline slicing strategies, i.e., 2D-axial, 2D-axial+, 2.5D, and 2.5D+, against our proposed *omnidirectional* volume slicing method introduced in Section 3.1. To ensure a fair comparison across all datasets and view configurations, each model employs an identical MLP pre-

diction head comprising 100k trainable parameters. The number of views is empirically set to 8, 16, and 24, providing a reasonable tradeoff between computational efficiency and slice diversity. The corresponding results are summarized in Table 1.

As visible from the results, the *2D-axial* approach performs worse than all alternative slicing methods, likely because the majority of volumetric information is discarded when relying on a single slice. We nevertheless included this experiment as it establishes a meaningful lower bound for the performance evaluation of different volume slicing methods. As indicated by the results, incorporating multiple axial slices represented by the *2D-axial+* approach improves performance by leveraging a greater amount of volumetric information.

Incorporating additional complementary perspectives, as in the *2.5D* slicing strategy, leads to a clear improvement in performance. For example, the average AUROC increases from 0.7335 for the *2D-axial+* configuration (three slices) to 0.7681 for the *2.5D* setup (three slices) across six different oncology datasets. This trend extends to the *2.5D+* variant, which consistently outperforms its *2D-axial+* counterpart for the same number of views. While the performance differences are more pronounced between 3, 8, and 16 views, the gap narrows at 24 views, suggesting that incorporating diverse perspectives is particularly beneficial when only a limited number of slices are available.

Beyond canonical plane slicing, our proposed *omnidirectional* slicing strategy, which integrates both canonical and non-canonical orientations, further improves downstream performance. On average, across all evaluated configurations, the *omnidirectional* approach consistently outperforms both the *2D-axial+* and *2.5D+* methods. Furthermore, for our volumetric slicing strategy, we observe a general trend across datasets. Increasing the number of views typically leads to higher performance. This relationship remains consistent for the *omnidirectional*

¹<http://github.com/compai-lab/2025-Media-kiechle>

²<https://pypi.org/project/OmniSlicer>

Table 1: Quantitative benchmarking of traditional volume slicing strategies (2D-axial, 2D-axial+, 2.5D, 2.5D+) against our proposed omnidirectional slicing strategy. All methods are evaluated for various numbers of volume slices, along with an MLP prediction head, using 5-fold cross-validation and AUROC as the performance metric. The rightmost column reports the average performance across datasets. The best-performing method on a per-dataset basis is indicated in bold.

Pred. Head	Volume Slicing Strategy	Views	Breast	Brain	Head-Neck	Kidney	Liver	Soft-Tissue	Mean
MLP	2D-axial	1	0.5844	0.9013	0.6192	0.6541	0.8821	0.7421	0.7305
		3	0.6166	0.8936	0.6550	0.6234	0.8803	0.7322	0.7335
		8	0.6402	0.9164	0.6171	<u>0.7689</u>	0.8738	0.7530	0.7615
		16	0.6467	0.9104	0.5982	0.7348	0.9053	0.7555	0.7584
		24	0.6299	0.9295	0.6277	0.7806	0.8895	0.7634	0.7701
	2.5D	3	0.6783	0.9062	0.6818	0.6815	<u>0.9662</u>	0.6951	0.7681
		8	<u>0.7002</u>	0.9296	0.6555	0.7300	0.9484	0.7488	0.7854
	2.5D+	16	0.7437	0.9226	0.6588	0.7606	0.9667	0.7473	0.7999
		24	0.6991	0.9265	0.6799	0.7041	0.9468	0.7751	0.7885
	Omnidirectional (ours)	8	0.6627	0.9457	0.6790	0.7306	0.9372	0.7702	0.7876
		16	0.6689	<u>0.9491</u>	<u>0.6999</u>	0.7450	0.9352	<u>0.8090</u>	<u>0.8012</u>
		24	0.6794	0.9492	0.7384	0.7229	0.9421	0.8602	0.8154

tional approach, with the sole exception of the kidney dataset, highlighting a clear positive correlation between the number of views and overall model performance. Furthermore, although the proposed *omnidirectional* slicing strategy does not achieve the highest performance on every individual dataset, it consistently surpasses its *2D-axial+* and *2.5D+* counterparts for the same number of views when considering the mean performance across all datasets. This demonstrates the method’s robust and, in several cases, superior capability in capturing complementary volumetric information relevant for downstream classification.

4.4. Performance Evaluation of TomoGraphView

Building on the benchmarking results presented in Section 4.3, our experiments demonstrate that the proposed *omnidirectional* slicing approach provides a robust and effective alternative to conventional volume slicing strategies, surpassing their capabilities across datasets and, remarkably, resulting in a higher overall average downstream performance. However, since the MLP prediction head is inadequate to account for relationships between omnidirectional slice encodings, we shift our focus toward aggregating slice-level representations using the proposed spherical mesh-based graph topology (see Section 3.3) and its corresponding graph-based feature aggregation mechanism. In this experiment, we compare *Omnidirectional Volume Slicing + MLP* (Section 4.3) against our proposed *TomoGraphView* framework (i.e., *Omnidirectional Volume Slicing + GNN*). Both methods operate on identical *omnidirectional* slice inputs, ensuring that any observed performance differences arise solely from the aggregation strategy rather than the input representations. The corresponding results are summarized in Table 2.

The key distinction lies in the GNN’s ability to explicitly integrate both local and global information by propagating slice-level features across nodes through edge connections defined by the graph structure (see Section 3.3). This mechanism is not

available to the MLP, which instead processes all slice information through fully connected layers without explicit relational modeling. Furthermore, given that the graph topology approximates a three-dimensional geometric structure (i.e., a sphere), we argue that a graph-based representation is better suited to preserve the inherent spatial relationships among slices compared to dense connections in an MLP. Consequently, the graph structure provides a more faithful representation of the 3D volume when represented through 2D slices and their corresponding relative position.

The advantage of propagating information across slices via a meaningful underlying graph structure is supported by the results in Table 2. Specifically, when comparing the average performance of the *Omnidirectional Volume Slicing + MLP* against our *TomoGraphView* framework (*Omnidirectional Volume Slicing + GNN*) across all six datasets, our proposed framework consistently outperforms the MLP for every view configuration, ultimately increasing from 0.8154 to 0.8372 in AUROC performance for the best performing model in each case. Notably, our *TomoGraphView* framework achieves superior performance across nearly all dataset–view combinations, with the exception of Head-Neck at 8 views and Soft-Tissue at 24 views. These findings strongly highlight the effectiveness of the graph-based approach over the MLP baseline.

4.5. Impact of Volume Slice Spacing

A closer examination of the results obtained with the proposed *TomoGraphView* framework, as shown in Table 2, reveals distinct dataset-specific trends. For head-neck, kidney, liver, and soft-tissue tumors, the best performance is achieved when using the highest number of views. In contrast, for breast and brain tumor datasets, configurations with 8 views yield superior results compared to those with a greater number of views. To better understand these patterns, we hypothesize that the observed performance differences with respect to the number of views are influenced by dataset-specific imaging characteris-

Table 2: Quantitative Area Under the Receiver Operating Characteristic Curve (AUROC) results for the performance comparison between omnidirectional volume slicing in combination with an MLP against our proposed TomoGraphView framework. Evaluation is performed across different datasets and view configurations. The rightmost column shows the average performance from each method and view. The best performing method on a per dataset basis is indicated in bold.

Method	Views	Breast	Brain	Head-Neck	Kidney	Liver	Soft-Tissue	Mean
Omnidirectional Volume Slicing + MLP	8	0.6627	0.9457	0.6790	0.7306	0.9372	0.7702	0.7876
	16	0.6689	0.9491	0.6999	0.7450	0.9352	0.8090	0.8012
	24	0.6794	0.9492	0.7384	0.7229	0.9421	0.8602	0.8154
Omnidirectional Volume Slicing + GNN TomoGraphView (ours)	8	0.7336	0.9678	0.6780	0.7878	0.9600	0.8399	0.8279
	16	0.6907	0.9666	0.7244	0.7479	0.9585	0.8389	0.8212
	24	0.7084	0.9655	0.7527	0.7893	0.9634	0.8441	0.8372

tics, particularly the voxel spacing along the axial dimension. As illustrated in Figure 5 (top), datasets with larger axial voxel spacing, such as soft-tissue, head-neck, and kidney, appear to benefit more from an increased number of views. This effect is reflected in the AUROC improvements observed when increasing the number of views from 8 to 24 in soft-tissue tumors (0.8399 vs. 0.8441), head and neck tumors (0.6780 vs. 0.7527), and kidney tumors (0.7878 vs. 0.7893).

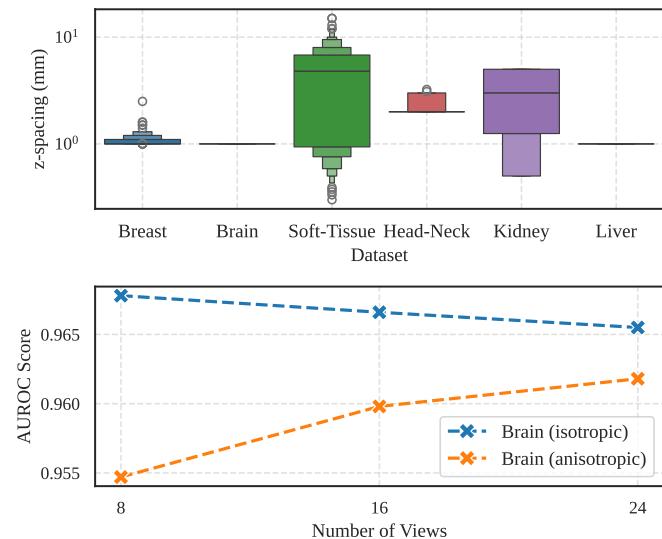


Figure 5: Box plot visualization of z-spacings (axial direction) across our utilized oncology datasets (top), and impact of volume slice spacings on the optimal number of views in terms of Area Under the Receiver Operating Characteristic Curve (AUROC) performance (bottom).

To further validate this empirical observation, we conducted an additional experiment by artificially increasing the z-spacing in a high-resolution dataset. For this purpose, we utilize the brain tumor dataset, which features isotropic $1 \times 1 \times 1$ mm spacing and provides a sufficiently large sample size ($n = 501$) for robust performance estimation. We resampled the isotropic samples to be anisotropic, specifically by increasing the axial spacing to 6 mm, while keeping the other dimensions unchanged, and reevaluated our proposed TomoGraphView framework for 8, 16, and 24 views.

We found two effects: First, the overall performance decreased from the isotropic to the anisotropic setting, which is

an expected consequence of information loss and resampling artifacts. Second, and more importantly, for the anisotropic setting, the performance across different view counts now follows the same trend as in high z-spacing datasets, such as soft-tissue, head-neck, or kidney, where increasing the number of views consistently improves results, as illustrated in Table 2. This finding suggests that additional views are more suitable for compensating for the loss of information introduced by anisotropic resampling.

4.6. Graph Topology Ablation

In this experiment, we conduct a graph topology ablation study, comparing different graph topologies to assess the impact of graph connectivity on our proposed *TomoGraphView* framework. In total, we compare five distinct graph topologies, which we refer to as *spherical*, *uniform*, *linear decay*, *inverse*, and *inverse-square*. More specifically, *uniform*, *linear decay*, *inverse*, and *inverse-square* topologies share the same graph connectivity (i.e., complete graph), but show differences in the edge-weights for node cross-connections (i.e., connections to non-immediate neighbors after triangulation as indicated in Section 3.3). The corresponding edge weights are determined according to the node distance in hops as indicated by Equation 3.3 and depicted in Figure 6 (top). In contrast, the *spherical* scheme does not use any node cross-connections, thus only connects nodes adjacent in triangulation (see Section 3.3), effectively representing a localized graph structure. Hence, the *spherical* topology results in a substantially sparser graph representation compared to the fully connected variants.

We conducted the graph topology experiments across all datasets and report the average AUROC performance using configurations with 24 nodes, as previous experiments in Section 4.4 demonstrated this to be the overall best-performing setting. Additionally, a larger number of nodes enables a more informative comparison, since the addition of edges has a greater relative effect in denser graphs. The corresponding results are presented in Figure 6 (bottom). Several key observations emerge from these experiments. First, the *spherical* topology achieves comparable performance to the fully connected *uniform* topology, despite relying on substantially fewer edges. This highlights the efficiency and relevance of the local spherical connectivity pattern. Second, introducing cross-connections between all nodes improves performance only when employing smooth distance-based weighting schemes such as *inverse*

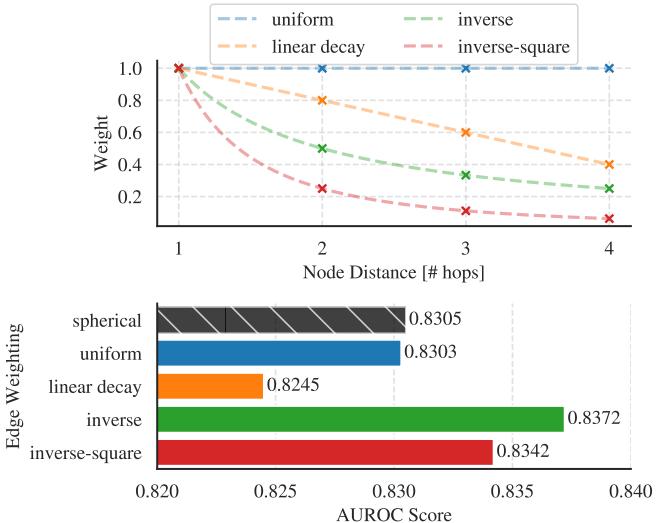


Figure 6: Graph Topology Analysis: Edge weighting schemes for uniform, inverse, inverse-square, and linear-decay according to the node hop distance (top). Graph topology results in terms of Area Under the Receiver Operating Characteristic Curve (AUROC) performance for each of the edge weighting schemes (bottom).

and *inverse-square*, which effectively expand the receptive field by aggregating information from diverse perspectives. Among these, the *inverse* weighting outperforms the *inverse-square* variant. In contrast, the *linear-decay* weighting scheme, although also expanding the receptive field, leads to a decline in AUROC performance. This finding highlights the importance of using decaying weighting functions that place a stronger relative emphasis on local connections (i.e., one-hop neighbors) compared to global ones (i.e., cross-connections). Detailed results, presented on a per-dataset basis for each of the different graph topologies, can be found in the appendix (Appendix D).

4.7. Slice-wise Feature Aggregation Methods

In this section, we compare *TomoGraphView* with two established slice-wise aggregation methods from the literature: the long short-term memory (LSTM) model, which interprets a 3D volume as a sequence of consecutive slice embeddings, and the medical slice transformer (MST), which employs an attention mechanism to fuse intra-slice information. To ensure a fair comparison, the number of trainable parameters is fixed to 100k across all models. Additionally, we evaluate both LSTM- and MST-based aggregation when combined with our proposed *omnidirectional slicing* strategy, thereby disentangling the effects of the aggregation mechanism from those of slice selection.

Extending the MST-based aggregation to *omnidirectional* slices does not require any explicit modeling of slice relationships, as the cross-attention mechanism inherently aims to capture inter-slice relationships. In contrast, this adaptation is less intuitive for LSTM models, which are inherently designed to process naturally sequential inputs. Since *omnidirectional* slices, derived from points sampled on a spherical surface, lack a natural ordering, we impose and fix a consistent sequence

based on node indices. The mean AUROC and MCC values comparing *TomoGraphView*, LSTM, and MST across all datasets are summarized in Table 3.

We start by evaluating the LSTM and MST approaches in their original formulation using consecutive axial slices (2D-axial+). The results show that the best-performing MST outperforms the best-performing LSTM for both evaluation metrics, with AUROC values of 0.7768 vs. 0.7446 and MCC values of 0.3995 vs. 0.3365, highlighting the strong aggregation capability of the MST. Furthermore, this performance advantage is consistent across all view configurations, with the MST surpassing the LSTM in both evaluation metrics. Nevertheless, both methods fall short of the proposed *TomoGraphView*, which achieves an AUROC of 0.8372 and an MCC of 0.5191.

Admittedly, comparing LSTM and MST in their original 2D-axial+ slicing configuration to *TomoGraphView* is only partially fair, as differences arise not only from the aggregation mechanism but also from the contextual information available in the input slices. To address this, we further extend the evaluation by applying *omnidirectional* slicing to both LSTM and MST. Two key observations emerge: (1) MST continues to outperform LSTM across all view configurations, and (2) both models experience substantial performance improvements. Specifically, for LSTM, *omnidirectional* slicing increases AUROC from 0.7446 to 0.7863 and MCC from 0.3365 to 0.4444. Similarly, for MST, AUROC improves from 0.7768 to 0.8198 and MCC from 0.3967 to 0.4732. These results highlight the clear benefit and broad applicability of the proposed *omnidirectional volume slicing* strategy. While both baselines show notable gains, *TomoGraphView* still achieves the highest overall performance.

A more detailed, dataset-wise comparison is presented in Figure 7, which reports AUROC (left) and MCC (right) scores for 24 views. Comprehensive numeric results across datasets, models, and view configurations are provided in Appendix Table E.5. Several trends can be observed: in terms of AUROC, *TomoGraphView* performs comparably to baseline methods on the breast, brain, and soft-tissue tumor datasets, but clearly surpasses them on the head-neck, kidney, and liver tumor datasets. A similar pattern holds for MCC, where *TomoGraphView* consistently outperforms all competing methods, except for the kidney tumor dataset, where both MST variants slightly exceed its performance. Overall, these results underscore the strong capability of *TomoGraphView* for slice-wise feature aggregation and establish it as a competitive alternative to existing approaches.

4.8. Comparative Analysis - 3D Pretrained Models

In the following, we compare our proposed *TomoGraphView* framework with large-scale pretrained 3D models. All compared models are convolution-based, as vision transformer architectures for 3D images are highly memory-intensive due to the substantially increased token length (i.e., number of patches) relative to 2D images. Moreover, relative to convolutional methods, vision transformers typically require much larger datasets for effective pretraining, while medical imaging data generally remains comparatively limited.

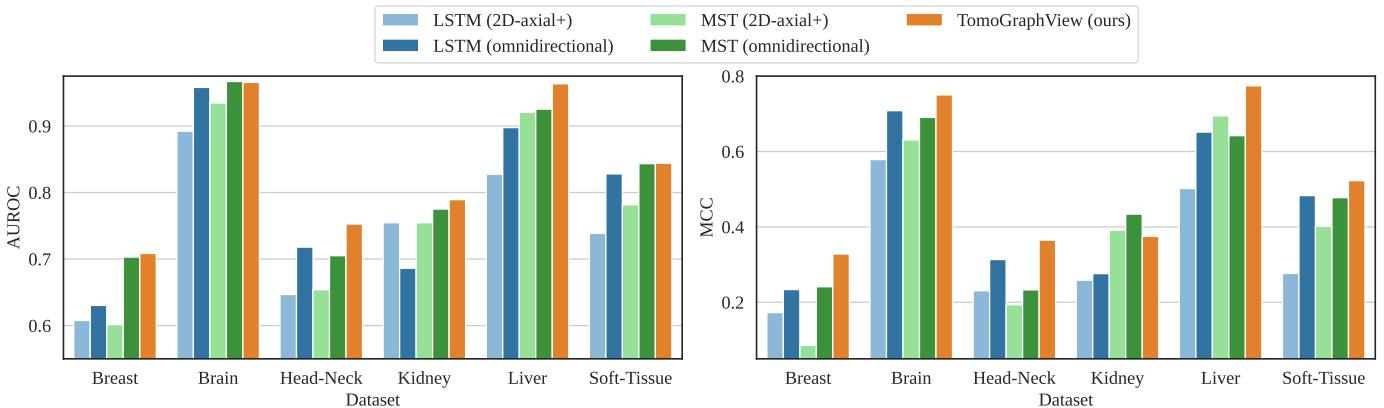


Figure 7: Detailed results comparing slice-wise aggregation methods, including our TomoGraphView framework, across datasets for Area Under the Receiver Operating Characteristic Curve (AUROC) and Matthews correlation coefficient (MCC) performance, indicated as the average for 5-folds.

Table 3: Quantitative results comparing slice-wise feature aggregation methods across their original slicing strategy (2D-axial+) and our proposed omnidirectional volume slicing, against our TomoGraphView framework. Here, LSTM corresponds to Long Short-Term Memory and MST to Medical Slice Transformer. Methods are compared across different numbers of views and performance metrics, such as the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Matthews correlation coefficient (MCC). Results indicate the average across all six datasets and 5-folds.

Model	Volume Slicing	Views	AUROC	MCC
LSTM	2D-axial+	8	0.7305	0.3101
		16	0.7330	0.3129
		24	0.7446	0.3365
LSTM	Omnidirectional	8	0.7764	0.4225
		16	0.7744	0.4178
		24	0.7863	0.4444
MST	2D-axial+	8	0.7583	0.3442
		16	0.7768	0.3967
		24	0.7746	0.3995
MST	Omnidirectional	8	0.8083	0.4691
		16	0.8171	0.4732
		24	0.8198	0.4530
TomoGraphView	Omnidirectional	8	0.8279	0.4891
		16	0.8212	0.5096
		24	0.8372	0.5191

The 3D models included in our comparison with *TomoGraphView* correspond to the top-performing approaches for downstream classification identified in the studies by Aerts et al. [1] and Wald et al. [76]. In particular, we benchmark our framework against FMCIB [51], Vista3D [27], Models Genesis [88], and two pretrained models introduced by Wald et al. [76] based on VoCo [81] and SwinUNETR [68]. A detailed description of the models is provided in Appendix C.

The remainder of this section is structured as follows. In Section 4.8.1, we first evaluate the fixed embeddings obtained from large-scale pretrained 3D models across all datasets, without applying any task-specific finetuning to assess their representational capacity learned during pretraining. Subsequently,

in Section 4.8.2, we perform dataset-specific finetuning, which typically improves downstream performance by allowing the pretrained representations to better adapt to the characteristics of each target task. Since all large-scale pretrained 3D models were trained in a self-supervised manner, we attach a classification head to the encoder to enable downstream classification. Consistent with our previous experiments, we fix the number of trainable parameters to 100k to ensure a fair comparison. Notably, prediction heads previously discussed, such as GNN, transformer, or LSTM-based architectures, cannot be integrated with the 3D pretrained models, as they require slice-based inputs rather than full 3D volume embeddings.

4.8.1. Frozen Backbones

We begin by evaluating the five large-scale pretrained 3D models across all six oncological downstream tasks and compare their performance with our proposed *TomoGraphView* framework. Results are depicted in the left radar plot of Figure 8. Detailed quantitative results across methods, datasets, and performance metrics are presented in Appendix Table F.6.

Across all tasks, the proposed *TomoGraphView* framework achieves an average AUROC of 0.8282, outperforming the on average second-best approach, FMCIB [51], which achieves an AUROC of 0.7170. This demonstrates a consistent advantage of *TomoGraphView* over all competing models. Looking more closely, on a per-task level, *TomoGraphView* also yields the highest AUROC for each dataset: breast tumors (0.7214 vs. 0.6207, SwinUNETR), brain tumors (0.9639 vs. 0.8985, FMCIB), head-neck tumors (0.6999 vs. 0.5910, SwinUNETR), kidney tumors (0.7750 vs. 0.7466, SwinUNETR), liver tumors (0.9664 vs. 0.8711, FMCIB), and soft-tissue tumors (0.8426 vs. 0.6926, SwinUNETR). Although the performance gains are less pronounced for kidney tumors (0.7750 vs. 0.7466) and brain tumors (0.9639 vs. 0.8985), *TomoGraphView* consistently outperforms all competing methods by a clear margin across the remaining breast, head-neck, liver, and soft-tissue tumor datasets. This observation, as exemplified by the AUROC results, can be further confirmed by additional evaluation metrics, such as balanced accuracy, F1-Score, and Matthews correlation coeffi-

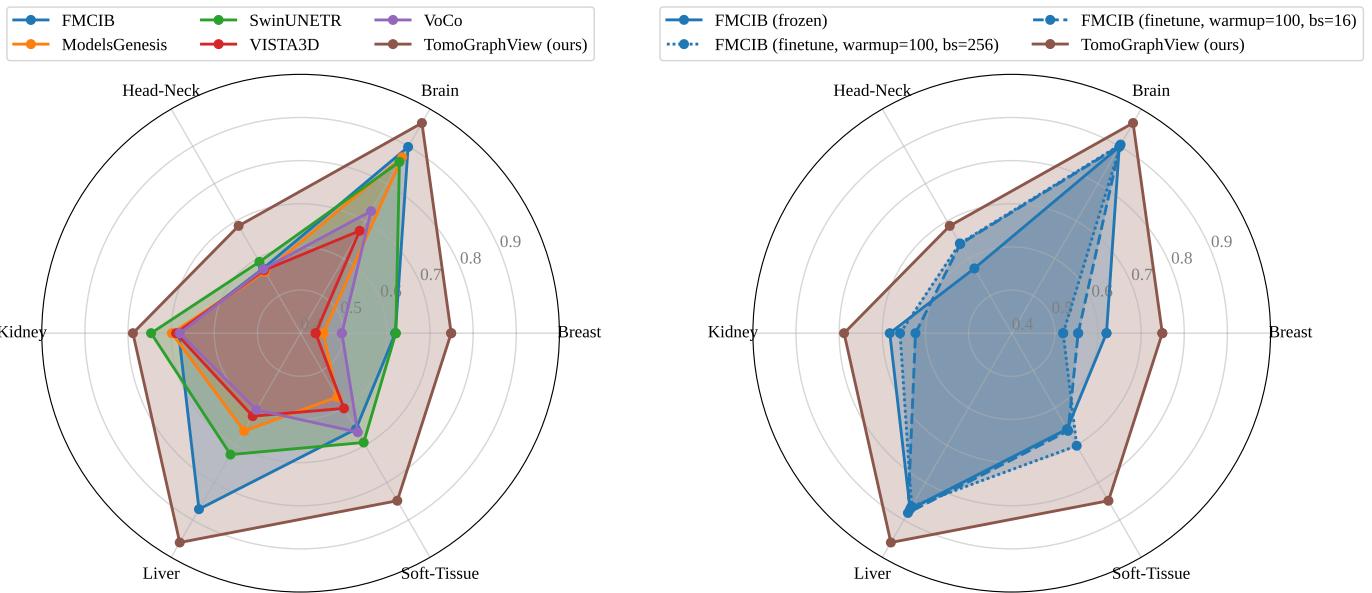


Figure 8: Radar plots showing quantitative results comparing frozen 3D backbones against our TomoGraphView framework (left) and fine-tuning results of the best performing 3D backbone (FMCIB) against our TomoGraphView framework (right). Plots show the average Area Under the Receiver Operating Characteristic Curve (AUROC) performance across 5-folds.

cient, underscoring the great potential of *TomoGraphView*. Detailed results can be found in the Appendix Table F.6.

4.8.2. 3D Model Fine-tuning Performance

Building on the frozen backbone results presented in the previous subsection, where FMCIB [51] emerged as the best-performing large-scale pretrained 3D model on average across all datasets in terms of AUROC, balanced accuracy, F1-score, and Matthews correlation coefficient, we now evaluate its fine-tuning performance, as it allows the pretrained representations to better adapt to the specific characteristics of each downstream task. Strictly following Wald et al. [76], we finetune FMCIB for 200 epochs in total, using a learning rate of 1×10^{-4} , a cosine annealing scheduler, a gradual warm-up of 100 epochs applied to both the encoder and the classification head, and a batch size of 256. Additionally, we explore a batch size of 16, which is consistent with previous experiments. The corresponding results are illustrated in the right radar plot of Figure 8, with detailed quantitative results provided in Appendix Table F.7, which additionally shows decreased warm-up durations of 20 and 50 epochs to allow for more update steps under higher learning rates.

Overall, fine-tuning with a 100-epoch warm-up and a batch size of 256 yields performance gains across several downstream tasks. In particular, AUROC improves for brain tumors from 0.8985 to 0.9049, soft-tissue tumors from 0.6567 to 0.7013, and head-neck tumors from 0.5733 to 0.6403. However, no measurable improvements are observed for the breast, liver, or kidney tumor datasets compared to the frozen FMCIB backbone. Given that the datasets showing no improvements are also those with relatively small sample sizes, we conduct an additional experiment using a reduced batch size of 16 to allow for more update steps during finetuning. This adjustment improves

AUROC performance for the liver tumor dataset from 0.8711 to 0.8811, whereas performance for the breast and kidney datasets remains inferior. These results suggest that the effectiveness of fine-tuning is highly dataset-dependent, and a single universal configuration is unlikely to generalize across datasets without extensive hyperparameter exploration. Notably, our proposed *TomoGraphView* framework continues to outperform the FMCIB model even under finetuning conditions.

5. Discussion

In this work, we introduce *TomoGraphView*, a novel framework designed for 3D medical image classification that preserves the spatial structure of volumetric medical imaging data while enabling compatibility with 2D feature encoders through both canonical and non-canonical 3D volume slicing and graph-based feature aggregation. We comprehensively evaluate our approach across six diverse oncology datasets covering multiple anatomical regions (e.g., brain, breast, head-neck, kidney, or liver), various imaging modalities (CT and multiple MRI sequences), and various classification tasks primarily concerning tumor characterization. Building on our two key contributions, *omnidirectional volume slicing* and *spherical mesh-based feature aggregation*, our experimental evaluations show that *TomoGraphView* achieves state-of-the-art results across downstream classification tasks, and even outperforms large-scale pretrained 3D models. The following section provides a detailed discussion of the results and findings obtained throughout this work.

5.1. Why Omnidirectional Volume Slicing Matters

As the first key contribution of this work, we introduce our novel *omnidirectional* volumetric slicing strategy, which extends beyond axial, coronal, and sagittal orientations, thereby

enabling a richer characterization of volumetric structures that are not naturally aligned with canonical axes. The value of rich characterization through both meaningful and representative volume slicing to describe target 3D structures is supported by our experimental evaluations in Section 3.1, which demonstrate that incorporating additional planes, transitioning from 2D axial to 2.5D slicing, consistently enhances downstream performance. On average, AUROC performance across all evaluation datasets improves by 4.71% (from 0.7335 to 0.7681) when comparing 2.5D slicing with its corresponding 3-view 2D-axial+ configuration. This trend also persists for a larger number of views: with 24 views, AUROC increases by 2.39% (from 0.7701 to 0.7885). However, critically, further gains are achieved with *omnidirectional* volume slicing, which improves AUROC by 3.41% (from 0.7885 to 0.8154) compared to 2.5D+ slicing, and by 5.88% (from 0.7701 to 0.8154) relative to pure axial slicing.

Notably, the benefit of *omnidirectional* volume slicing generalizes beyond MLPs across various prediction head backbones, including LSTM, and transformer-based architectures, as outlined in Section 4.7. Interestingly, although our *omnidirectional* slicing strategy cannot be directly converted into a naturally meaningful slice ordering, as compared to consecutive axial slicing, LSTMs trained on *omnidirectional* slices outperform those trained on consecutive axial slices, with a notable AUROC increase of 5.60% (from 0.7446 to 0.7863). Similarly, transformer-based models show an AUROC improvement of 5.83% (from 0.7746 to 0.8198).

Overall, the superior results obtained throughout this work demonstrate that *omnidirectional* slicing provides a richer representation of volumetric structures, yielding consistent and architecture-agnostic performance enhancements. These findings highlight the versatility and added value of our proposed slicing strategy as a compelling alternative to traditional approaches.

5.2. Graph-based Feature Aggregation Excels Baselines

Beyond the choice of representing a volume as a set of 2D slices, where *omnidirectional* volume slicing emerged as the primary choice throughout our experiments, the way information from cross-sectional images is aggregated is of central importance. To this end, we present our second core contribution in this work, GNN-based feature aggregation, which relies on a *spherical mesh-based graph topology* to explicitly model the spatial relationships of encoded 2D slices extracted from 3D volumes. Our spherical graph design is derived from the *omnidirectional* slicing strategy, enabling the integration of information across a set of in-plane and out-of-plane perspectives while explicitly encoding spatial relationships within the graph topology. Our experimental evaluations in Section 4.4 and Section 4.7 underscore the advantage of our graph-based feature aggregation method across datasets and backbones.

While MLPs do not offer any built-in capability to model relationships among slices, LSTMs treat consecutive slices as temporal sequences, thereby preserving their spatial ordering within a volume. However, such an ordering is only naturally defined along a single plane of consecutive slices, restricting

the representation to one direction and potentially overlooking complementary information from other orientations. Nevertheless, although *omnidirectional* volume slicing improves LSTM downstream performance, there is no natural ordering of these slices, potentially hindering further improvements.

We attribute the superior performance of our GNN-based approach to its stronger inductive bias. Specifically, the node feature aggregation and node connectivity are predefined, whereas for transformers, such as MST-based aggregation, spatial relationships and global dependencies must be learned through their self-attention mechanism, introducing additional complexity. Moreover, transformers typically perform best in large-scale data settings, which are uncommon in medical imaging, limiting their practical suitability. We argue that simplifying the learning task by the topological constraints introduced in our GNN-based approach leads to improved overall average performance while handling limited data more efficiently.

5.3. Dimensionality-Quantity Tradeoff in 3D Classification

Current approaches to 3D medical image classification face a fundamental trade-off between data dimensionality and data quantity. Although 2D slice-based methods achieve strong performance, they inherently lose anatomical context and spatial coherence when 3D volumes are decomposed into individual slices. Conversely, models that directly process 3D data preserve spatial structure but are constrained by the limited availability of pretraining data. Both dimensionality and data scarcity substantially impact a model’s representational capacity for volumetric analysis, a crucial aspect for reliable downstream classification.

While foundation models have advanced rapidly in domains such as histopathology [23, 12, 38], where large-scale 2D data and sophisticated pretraining strategies drive strong transferability, native 3D foundation models remain comparatively underdeveloped [74]. Our findings in Section 4.8 support this observation.

We begin by comparing *TomoGraphView* with five large-scale pretrained 3D models across six oncology datasets, using frozen encoder backbones. Evaluating raw embeddings enables a resource-efficient and faithful assessment of the representational quality learned during pretraining. Ideally, a robust pretrained model should produce embeddings that capture rich image semantics and correlate well with downstream performance. As *TomoGraphView* also relies solely on frozen DINOv2 features, this comparison remains fair. Across all datasets, the FMCIB model [51] performs best among 3D approaches but consistently trails *TomoGraphView*, with relative improvements of 15.5% in AUROC and 16.8% in balanced accuracy (see Table F.6 in the Appendix).

Interestingly, despite being pretrained exclusively on body CT data (DeepLesion [84]), FMCIB even outperforms MRI-pretrained models on brain tumor MRI classification and matches the performance of the MRI-pretrained Swin-UNETR [76]. We attribute this cross-modality generalization to the model’s ability to learn universal imaging features such as textures, edges, and spatial hierarchies.

We further evaluate fine-tuning performance to assess how well FMCIB adapts pretrained representations to specific downstream tasks. Following the fine-tuning setup of Wald et al. [76], we observe improvements for brain, soft-tissue, and head-neck tumors, but no gain for breast, kidney, or liver datasets. This highlights that the effectiveness of fine-tuning is highly dataset-dependent and that no single configuration generalizes universally without extensive hyperparameter optimization. Nevertheless, overall fine-tuning performance still falls short of *TomoGraphView*. In this study, we focused exclusively on finetuning large-scale pretrained 3D models, as transfer learning has consistently demonstrated superior performance over training from scratch, particularly in medical image classification tasks where labeled data is often scarce. Existing studies show that leveraging prior knowledge from large-scale datasets enables faster convergence and improved generalization to domain-specific medical data [66].

In summary, the results from Sections 4.8.1 and 4.8.2 highlight the added value of combining 2D foundation models with advanced volume-slicing and feature-aggregation strategies. We therefore position *TomoGraphView* as a key step toward bridging the gap until fully native 3D foundation models become available for medical image analysis.

5.4. Limitations and Outlook

While our proposed omnidirectional volume slicing consistently improves performance across tasks, datasets, and prediction heads, it currently depends on the availability of tumor segmentation masks, as slice selection is based on identifying the largest lesion area. Notably, all other volumetric slicing methods in this study also rely on segmentation masks for this purpose. Fortunately, many established algorithms can generate sufficiently accurate tumor delineations, making our approach broadly applicable. Nevertheless, it is worth noting that 3D medical image classification is usually performed on subvolumes, typically defined by bounding box coordinates around the structure of interest, to reduce the complexity of the problem. Given such a bounding box, we argue that selecting the central slice of the resulting subvolume should provide adequate information, thereby bypassing the need for precise tumor segmentation. Moreover, adding transparency, our *omnidirectional* volume slicing strategy incurs a modest computational overhead during preprocessing, as generating N omnidirectional slices requires interpolating the volume N times from different orientations. To address current limitations, future work will evaluate our framework on datasets providing bounding box annotations instead of segmentation masks. Furthermore, we plan to extend *TomoGraphView* by integrating a detection module, enabling a fully end-to-end 3D medical image classification pipeline. A promising direction in this regard is leveraging multimodal large language models, which can utilize in-context samples and textual prompts to predict bounding boxes for relevant structures, without requiring costly parameter updates or finetuning [85]. Furthermore, as *TomoGraphView* is designed as a flexible framework compatible with any 2D foundation model, we anticipate that its performance will con-

tinue to improve in tandem with future advancements in 2D backbone architectures and training paradigms.

6. Conclusion

In this work, we present *TomoGraphView*, a unified framework for 3D medical image classification that combines *omnidirectional* volume slicing with *spherical graph-based feature aggregation*. Using six diverse datasets spanning different anatomical regions, imaging modalities, and classification tasks, we demonstrated that (I) our omnidirectional slicing approach consistently outperforms standard slicing strategies based on axial, sagittal, coronal, or combined views, and (II) *spherical graph-based feature aggregation* surpasses the performance of existing slice-wise feature aggregation baselines.

Although three-dimensional models hold great promise in medical image analysis, particularly for 3D tomography data, as they leverage spatial relationships across all axes to better capture anatomical context, their adoption in practice is impeded by several challenges: They are computationally demanding, require substantially larger datasets to achieve robust generalization due to their high parameter counts, and offer limited possibilities for transfer learning, since publicly available pretrained 3D networks remain comparatively scarce and less established than 2D alternatives [74]. As a consequence, slice-based 2D approaches remain an attractive choice, as they are more efficient to train, naturally benefit from decomposing volumetric scans into multiple slices, and can exploit powerful pretrained backbones developed for natural image tasks.

This is where *TomoGraphView* comes into the picture. We promote our presented *omnidirectional* volume slicing strategy as an attractive alternative to existing, canonical volume slicing methods, as it has been shown to improve downstream performance across a plethora of prediction heads, including MLP, LSTM, and transformer-based approaches. Moreover, by integrating *omnidirectional* volume slicing with spherical graph-based feature aggregation, our *TomoGraphView* framework achieves superior performance levels when compared against both existing slice-wise feature aggregation methods and large-scale pretrained 3D models. The experimental evaluations with their corresponding results in this work underscore the added value of combining 2D foundation models with advanced volume slicing strategies and feature-aggregation techniques, and therefore represent a key technology toward bridging the gap until fully native 3D foundation models become available for medical image analysis, thereby highlighting the need of our proposed approach to enable more accurate, personalized, and scalable solution for tomography-based image diagnostics.

In summary, our presented work highlights the versatility and promise of 2D slice-wise strategies for 3D medical image classification. We believe that combining 2D foundation models with advanced volumetric slicing techniques and graph-based slice-wise feature aggregation methods represents a key step toward bridging the gap until fully native 3D foundation models become available in medical image analysis.

Acknowledgments

Authors JK, SMF, and JAS gratefully acknowledge financial funding received by Munich Center for Machine Learning, sponsored by the German Federal Ministry of Education and Research. Authors JCP and JK gratefully acknowledge funding from the Wilhelm Sander Foundation in Cancer Research (2022.032.1). Author SMF has received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 515279324 / SPP 2177.

Ethical Standards

Ethical clearance for this study utilizing retrospectively collected in-house soft-tissue sarcoma data was obtained from the respective ethics committees of the data-providing institutions (i.e., Technical University of Munich, University of Washington/Seattle Cancer Care Alliance).

Data Availability

All data, except the in-house soft-tissue sarcoma dataset, used in this study, are publicly available. We provide detailed descriptions and proper citations for each dataset used to ensure full transparency and reproducibility. Additionally, all code for data preprocessing and preparation is publicly released and can be used to fully reproduce our results. The code is available at: <http://github.com/compai-lab/2025-MedIA-kiechle>. The in-house soft-tissue sarcoma data may be obtained from the corresponding author on reasonable request and may be published online in the future after approval by the institutional ethics review boards. We publicly share our accessible code base at <http://github.com/compai-lab/2025-MedIA-kiechle> and provide a user-friendly library for omnidirectional volume slicing at <https://pypi.org/project/OmniSlicer>.

Conflicts of Interest

JCP holds shares in Mevidence and has received honoraria from AstraZeneca and Brainlab, all of which are outside the scope of the submitted work. The remaining authors declare that they have no known conflict of interest nor competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of AI Assistance

The authors acknowledge the use of ChatGPT to assist in improving the readability and language of this manuscript. The authors thoroughly reviewed and edited the text following its use and accept full responsibility for the content and integrity of the final publication.

Authorship Contribution Statement

Johannes Kiechle: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stefan M. Fischer:** Validation, Writing - Review & Editing. **Daniel M. Lang:** Validation, Writing - Review & Editing. **Cosmin I. Bercea:** Validation, Writing - Review & Editing. **Matthew J. Nyflot:** Data, Validation, Writing - Review & Editing. **Lina Felsner:** Validation, Writing - Review & Editing. **Julia A. Schnabel:** Validation, Writing - Review & Editing, Supervision. **Jan C. Peeken:** Validation, Writing - Review & Editing, Supervision, Funding acquisition

References

- [1] Aerts, H., Pai, S., Hadzic, I., Fedorov, A., Mak, R., 2025. Foundation model embeddings for quantitative tumor imaging biomarkers. Research Square , rs-3.
- [2] Ahamed, S., Xu, Y., Bloise, I., HO, J., Uribe, C.F., Dodhia, R., Ferres, J.L., Rahmim, A., 2023. A slice classification neural network for automated classification of axial pet/ct slices from a multi-centric lymphoma dataset, in: Medical Imaging 2023: Image Processing, SPIE. pp. 406–415.
- [3] Antropova, N., Abe, H., Giger, M.L., 2018. Use of clinical mri maximum intensity projections for improved breast lesion classification with deep convolutional neural networks. Journal of Medical Imaging 5, 014503–014503.
- [4] Ayzenberg, L., Giryes, R., Greenspan, H., 2024. Dinov2 based self supervised learning for few shot medical image segmentation, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.
- [5] Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerooni, A., Rekik, I., Merhof, D., 2023. Foundational models in medical imaging: A comprehensive survey and future vision. arXiv preprint arXiv:2310.18689 .
- [6] Baharoon, M., Qureshi, W., Ouyang, J., Xu, Y., Aljouie, A., Peng, W., 2024. Evaluating general purpose vision foundation models for medical image analysis: An experimental study of dinov2 on radiology benchmarks. URL: <https://arxiv.org/abs/2312.02366>, arXiv:2312.02366.
- [7] Bronstein, M.M., Bruna, J., Cohen, T., Veličković, P., 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478 .
- [8] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine 34, 18–42.

- [9] Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P., 2022. The university of california san francisco preoperative diffuse glioma mri dataset. *Radiology: Artificial Intelligence* 4, e220058.
- [10] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2022. Swin-unet: Unet-like pure transformer for medical image segmentation, in: European conference on computer vision, Springer. pp. 205–218.
- [11] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9650–9660.
- [12] Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al., 2024. Towards a general-purpose foundation model for computational pathology. *Nature medicine* 30, 850–862.
- [13] Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. Simclr: A simple framework for contrastive learning of visual representations, in: International Conference on Learning Representations, PMLR New York, NY, USA.
- [14] Corso, G., Cavalleri, L., Beaini, D., Liò, P., Veličković, P., 2020. Principal neighbourhood aggregation for graph nets. *Advances in neural information processing systems* 33, 13260–13271.
- [15] Cui, B., Islam, M., Bai, L., Ren, H., 2024. Surgical-dino: adapter learning of foundation models for depth estimation in endoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery* 19, 1013–1020.
- [16] Dadoun, H., Rousseau, A.L., de Kerviler, E., Correas, J.M., Tissier, A.M., Joujou, F., Bodard, S., Khezzane, K., de Margerie-Mellon, C., Delingette, H., et al., 2022. Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal us images. *Radiology: Artificial Intelligence* 4, e210110.
- [17] Delaunay, B., 1934. Sur la sphère vide. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskii i Estestvennyka Nauk* 7, 793–800.
- [18] Di Piazza, T., Lazarus, C., Nempong, O., Boussel, L., 2025. Structured spectral graph learning for anomaly classification in 3d chest ct scans. arXiv preprint arXiv:2508.01045 .
- [19] Dijkstra, E.W., 2022. A note on two problems in connexion with graphs, in: Edsger Wybe Dijkstra: his life, work, and legacy, pp. 287–290.
- [20] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.
- [21] Erdur, A.C., Rusche, D., Scholz, D., Kiechle, J., Fischer, S., Llorian-Salvador, O., Buchner, J.A., Nguyen, M.Q., Etzel, L., Weidner, J., et al., 2025. Deep learning for autosegmentation for radiotherapy treatment planning: State-of-the-art and novel perspectives. *Strahlentherapie und Onkologie* 201, 236–254.
- [22] Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y., 2024. Eva-02: A visual representation for neon genesis. *Image and Vision Computing* 149, 105171.
- [23] Filiot, A., Jacob, P., Mac Kain, A., Saillard, C., 2024. Phikon-v2, a large and public feature extractor for biomarker prediction. arXiv preprint arXiv:2409.09173 .
- [24] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E., 2017. Neural message passing for quantum chemistry, in: International conference on machine learning, Pmlr. pp. 1263–1272.
- [25] Grossberg, A.J., Mohamed, A.S., Elhalawani, H., Bennett, W.C., Smith, K.E., Nolan, T.S., Williams, B., Chamchod, S., Heukelom, J., Kantor, M.E., et al., 2018. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Scientific data* 5, 1–10.
- [26] Hamilton, W., Ying, Z., Leskovec, J., 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30.
- [27] He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., et al., 2025. Vista3d: A unified segmentation foundation model for 3d medical imaging, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 20863–20873.
- [28] Heller, N., Sathianathan, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445 .
- [29] Herent, P., Schmauch, B., Jehanno, P., Dehaene, O., Saillard, C., Balleyguier, C., Arfi-Rouche, J., Jégou, S., 2019. Detection and characterization of mri breast lesions using deep learning. *Diagnostic and interventional imaging* 100, 219–225.
- [30] Hu, Q., Whitney, H.M., Li, H., Ji, Y., Liu, P., Giger, M.L., 2021. Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection mri in breast cancer diagnosis using dynamic

contrast-enhanced mri. Radiology: Artificial Intelligence 3, e200159.

- [31] Huang, Y., Zou, J., Meng, L., Yue, X., Zhao, Q., Li, J., Song, C., Jimenez, G., Li, S., Fu, G., 2024. Comparative analysis of imagenet pre-trained deep learning models and dinov2 in medical imaging classification, in: 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), IEEE. pp. 297–305.
- [32] Huix, J.P., Ganeshan, A.R., Haslum, J.F., Söderberg, M., Matsoukas, C., Smith, K., 2024. Are natural domain foundation models useful for medical image classification?, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 7634–7643.
- [33] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.
- [34] Jang, J., Hwang, D., 2022. M3t: three-dimensional medical image classifier using multi-plane and multi-slice transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 20718–20729.
- [35] Keriven, N., 2022. Not too little, not too much: a theoretical analysis of graph (over) smoothing. Advances in Neural Information Processing Systems 35, 2268–2281.
- [36] Kiechle, J., Lang, D.M., Fischer, S.M., Felsner, L., Peeken, J.C., Schnabel, J.A., 2024. Graph neural networks: A suitable alternative to mlps in latent 3d medical image classification?, in: International Workshop on Graphs in Biomedical Image Analysis, Springer. pp. 12–22.
- [37] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015–4026.
- [38] Koch, V., Wagner, S.J., Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J.A., Peng, T., Marr, C., 2024. Dinobloom: a foundation model for generalizable cell embeddings in hematology, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 520–530.
- [39] Kraus, K.M., Oreshko, M., Bernhardt, D., Combs, S.E., Peeken, J.C., 2023. Dosiomics and radiomics to predict pneumonitis after thoracic stereotactic body radiotherapy and immune checkpoint inhibition. Frontiers in Oncology 13, 1124592.
- [40] Kwan, J., Su, J., Huang, S., et al., 2019. Data from radiomic biomarkers to refine risk models for distant metastasis in oropharyngeal carcinoma. the cancer imaging archive, 2019. ^ editors. Book Data from radiomic biomarkers to refine risk models for distant metastasis in oropharyngeal carcinoma, the cancer imaging archive .
- [41] Lilhore, U.K., Sharma, Y.K., Shukla, B.K., Vadlamudi, M.N., Simaiya, S., Alroobaea, R., Alsafyani, M., Baqasah, A.M., 2025. Hybrid convolutional neural network and bi-lstm model with efficientnet-b0 for high-accuracy breast cancer detection and classification. Scientific Reports 15, 12082.
- [42] Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Zhang, Y., Li, J., Bercea, C.I., Ouyang, C., Chen, C., Xiong, Z., Wiestler, B., Wachinger, C., Rueckert, D., Bai, W., Arcucci, R., 2025. Does dinov3 set a new medical vision standard? URL: <https://arxiv.org/abs/2509.06467>, arXiv:2509.06467.
- [43] Luo, J., Wan, X., Du, J., Liu, L., Zhao, L., Peng, X., Wu, M., Huang, S., Nie, X., 2025. Comprehensive multi-phase 3d contrast-enhanced ct imaging for primary liver cancer. Scientific Data 12, 768.
- [44] Meneghetti, A.R., Hernández, M.L., Kuehn, J.P., Löck, S., Carrero, Z.I., Perez-Lopez, R., Bressem, K.K., Brinker, T.J., Pearson, A.T., Truhn, D., et al., 2025. End-to-end prediction of clinical outcomes in head and neck squamous cell carcinoma with foundation model-based multiple instance learning. BMC Artificial Intelligence 1, 3.
- [45] Miron, R., Moisii, C., Dinu, S., Breaban, M.E., 2021. Evaluating volumetric and slice-based approaches for covid-19 detection in chest cts, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 529–536.
- [46] Müller-Franzes, G., Khader, F., Siepmann, R., Han, T., Kather, J.N., Nebelung, S., Truhn, D., 2025. Medical slice transformer for improved diagnosis and explainability on 3d medical images with dinov2. Scientific Reports 15, 23979.
- [47] Naser, M.A., Deen, M.J., 2020. Brain tumor segmentation and grading of lower-grade glioma using deep learning in mri images. Computers in biology and medicine 121, 103758.
- [48] Navarro, F., Dapper, H., Asadpour, R., Knebel, C., Spraker, M.B., Schwarze, V., Schaub, S.K., Mayr, N.A., Specht, K., Woodruff, H.C., et al., 2021. Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using mr imaging. Cancers 13, 2866.
- [49] Nguyen, N.T., Tran, D.Q., Nguyen, N.T., Nguyen, H.Q., 2020. A cnn-lstm architecture for detection of intracranial hemorrhage on ct scans. medRxiv , 2020–04.
- [50] Quab, M., Darzet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa,

- F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 .
- [51] Pai, S., Bontempi, D., Hadzic, I., Prudente, V., Sokač, M., Chaunzwa, T.L., Bernatz, S., Hosny, A., Mak, R.H., Birkbak, N.J., et al., 2024. Foundation model for cancer imaging biomarkers. *Nature machine intelligence* 6, 354–367.
- [52] Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A., 2025. Foundation models in radiology: What, how, why, and why not. *Radiology* 314, e240597.
- [53] Perez-Garcia, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., et al., 2025. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence* 7, 119–130.
- [54] Prakash, S.S., Visakha, K., 2020. Breast cancer malignancy prediction using deep learning neural networks, in: 2020 second international conference on inventive research in computing applications (ICIRCA), IEEE. pp. 88–92.
- [55] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.
- [56] Rana, M., Bhushan, M., 2023. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications* 82, 26731–26769.
- [57] Rasuli, F., 2025. Deep learning in radiology: Challenges and opportunities for improved diagnosis, in: 2025 International Conference on Frontier Technologies and Solutions (ICFTS), IEEE. pp. 1–10.
- [58] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.
- [59] Saha, A., Harowicz, M., Grimm, L., Weng, J., Cain, E., Kim, C., Ghate, S., Walsh, R., Mazurowski, M., 2021. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. the cancer imaging archive. Bethesda, MD, USA .
- [60] Saint-Esteven, A.L.G., Bogowicz, M., Konukoglu, E., Rieser, O., Balempas, P., Guckenberger, M., Tanadini-Lang, S., van Timmeren, J.E., 2022. A 2.5 d convolutional neural network for hpv prediction in advanced oropharyngeal cancer. *Computers in biology and medicine* 142, 105215.
- [61] Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 61–80.
- [62] Schulz, S., Woerl, A.C., Jungmann, F., Glasner, C., Stenzel, P., Strobl, S., Fernandez, A., Wagner, D.C., Haferkamp, A., Mildnerger, P., et al., 2021. Multimodal deep learning for prognosis prediction in renal cancer. *Frontiers in oncology* 11, 788740.
- [63] Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al., 2025. Dinov3. arXiv preprint arXiv:2508.10104 .
- [64] Song, X., Xu, X., Yan, P., 2024. Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 608–617.
- [65] Spohn, S.K., Schmidt-Hegemann, N.S., Ruf, J., Mix, M., Benndorf, M., Bamberg, F., Makowski, M.R., Kirste, S., Rühle, A., Nouvel, J., et al., 2023. Development of psma-pet-guided ct-based radiomic signature to predict biochemical recurrence after salvage radiotherapy. *European journal of nuclear medicine and molecular imaging* 50, 2537–2547.
- [66] Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* 35, 1299–1312.
- [67] Takahashi, K., Fujioka, T., Oyama, J., Mori, M., Yamaga, E., Yashima, Y., Imokawa, T., Hayashi, A., Kujiraoka, Y., Tsuchiya, J., et al., 2022. Deep learning using multiple degrees of maximum-intensity projection for pet/ct image classification in breast cancer. *Tomography* 8, 131–141.
- [68] Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 20730–20740.
- [69] Thomson, J.J., 1904. Xxiv. on the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 7, 237–265.
- [70] Thrall, J.H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K., Brink, J., 2018. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology* 15, 504–508.

- [71] Truong, T., Mohammadi, S., Lenga, M., 2021. How transferable are self-supervised features in medical image classification tasks?, in: Machine Learning for Health, PMLR. pp. 54–74.
- [72] Tschanne, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al., 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 .
- [73] Veasey, B.P., Amini, A.A., 2024. Parameter-efficient fine-tuning of dinov2 vision transformers for lung nodule classification, in: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1–5.
- [74] van Veldhuizen, V., Botha, V., Lu, C., Cesur, M.E., Lipman, K.G., de Jong, E.D., Horlings, H., Sanchez, C.I., Snoek, C.G., Wessels, L., et al., 2025. Foundation models in medical imaging—a review and outlook. arXiv preprint arXiv:2506.09095 .
- [75] Venkataraman, S., Pariza, V., Salehi, M., Knobel, L., Gidaris, S., Ramzi, E., Bursuc, A., Asano, Y.M., 2025. Franca: Nested matryoshka clustering for scalable visual representation learning. arXiv preprint arXiv:2507.14137 .
- [76] Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Köhler, G., Maier-Hein, K.H., 2024. An openmind for 3d medical vision self-supervised learning. arXiv preprint arXiv:2412.17041 .
- [77] Wang, R., Gao, L., Zhang, X., Han, J., 2024. Svfr: A novel slice-to-volume feature representation framework using deep neural networks and a clustering model for the diagnosis of alzheimer’s disease. *Heliyon* 10.
- [78] Wee, L., Dekker, A., 2019. Data from head-neck-radiomics-hn1. (No Title) .
- [79] Wei, X., Yu, R., Sun, J., 2020. View-gcn: View-based graph convolutional network for 3d shape analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
- [80] Wei, X., Yu, R., Sun, J., 2023. Learning view-based graph convolutional network for multi-view 3d shape analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7525–7541. doi:10.1109/TPAMI.2022.3221785.
- [81] Wu, L., Zhuang, J., Chen, H., 2024. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 22873–22882.
- [82] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems* 32, 4–24.
- [83] Xu, K., Hu, W., Leskovec, J., Jegelka, S., 2018. How powerful are graph neural networks? arXiv preprint arXiv:1810.00826 .
- [84] Yan, K., Wang, X., Lu, L., Summers, R.M., 2018. Deeplession: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging* 5, 036501–036501.
- [85] Ye, J., Tang, H., 2025. Multimodal large language models for medicine: A comprehensive survey. arXiv preprint arXiv:2504.21051 .
- [86] Yin, Y., Yakar, D., Slangen, J.J., Hoogwater, F.J., Kwee, T.C., de Haas, R.J., 2023. The value of deep learning in gallbladder lesion characterization. *Diagnostics* 13, 704.
- [87] Zhang, H., Zhang, J., Zhang, Q., Kim, J., Zhang, S., Gauthier, S.A., Spincemaille, P., Nguyen, T.D., Sabuncu, M., Wang, Y., 2019. Rsanet: recurrent slice-wise attention network for multiple sclerosis lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 411–419.
- [88] Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 384–393.
- [89] Zhu, W., Xie, L., Han, J., Guo, X., 2020. The application of deep learning in cancer prognosis prediction. *Cancers* 12, 603.
- [90] Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J., 2023. Segment everything everywhere all at once. *Advances in neural information processing systems* 36, 19769–19782.

Appendix A. Volume Slicing Comparison

While Figure 2 in Section 3.1 shows the resulting eight cross-sectional slices for a synthetically generated input volume, which displays a target structure that is not naturally aligned with canonical axes, we extend the visualization with 24 views in Figure A.9 below. While the slice variability for both, 2D-axial+ (top) and 2.5D+ (middle), is limited, our *omnidirectional* volume slicing strategy (bottom) excels by representing the target structure with all its facets, thereby depicting a more faithful representation of the target 3D structure.

Appendix B. Dataset Task Descriptions

For all datasets except the head-neck and liver cohorts, this study focuses on tumor grading, where cases are categorized into entities of differing malignancy. Generally, tumor grade provides critical insight into how aggressive a tumor is and its potential for progression or metastasis, thereby serving as an essential factor for cancer staging and treatment decisions. For the liver dataset, we perform tumor subtype prediction, whereas for the Head-Neck dataset, we instead predict human papillomavirus (HPV) status, as HPV infection constitutes a major risk factor for oropharyngeal cancer (OPC). Importantly, HPV-positive OPCs have been shown to exhibit greater radio-sensitivity and lower cancer-related mortality compared to HPV-negative cases. Thus, determining HPV status constitutes a key diagnostic marker with direct implications for prognosis and treatment decisions.

Appendix C. 3D Pretrained Model Description

Models Genesis [88] is a collection of models built from unlabeled 3D imaging data using a restorative reconstruction-based self-supervised method. It aims to generate powerful application-specific target models through transfer learning. The models demonstrate strong performance across various medical imaging tasks, emphasizing their potential for broad applicability in clinical settings.

VOCO [81] is a large-scale 3D medical image pre-training framework that leverages geometric context priors to learn consistent semantic representations. It is built on a substantial dataset of CT volumes and employs a novel pretext task for contextual position predictions. VOCO has demonstrated superior performance across various downstream tasks, establishing itself as a leading model in the field of medical imaging.

VISTA3D [27] is a versatile imaging segmentation and annotation model that supports both automatic and interactive segmentation of 3D medical images. It is the first unified foundation model to achieve state-of-the-art performance across 127 classes and is designed to facilitate efficient human correction through its interactive features. VISTA3D integrates a novel supervoxel method to enhance zero-shot performance, marking a significant advancement in 3D medical imaging.

FMCIB [51] is a foundation model designed to distinguish between lesions and non-lesions at the patch level in medical

imaging. It aims to enhance the detection and characterization of cancerous lesions by leveraging self-supervised learning techniques. It is based on SimCLR [13] (Simple Framework for Contrastive Learning of Visual Representations), a self-supervised learning approach that utilizes contrastive learning to pretrain deep neural networks without the need for labeled data. The core idea behind SimCLR is to maximize the similarity between differently augmented views of the same image while minimizing the similarity between views of different images.

SwinUNETR [10] was proposed as a pretraining method for the identically named SwinUNETR architecture and is composed of three components: image inpainting, Rotation prediction, and a contrastive training objective. The inpainting itself is a simple L1 loss applied to a masked-out image region, the rotation is a rotation of 0°, 90°, 180°, or 270° degrees along the z-axis, with an MLP used to classify the applied rotation. Lastly, the contrastive coding ensures that the linearly projected representations of the encoder are highly similar or dissimilar, depending on whether the two sub-volumes belong to the same or a different image, respectively. These three losses are combined with an equal weighting to form the SwinUNETR pre-training.

Appendix D. Graph Topology Ablation

Detailed quantitative results of our graph topology ablation experiment performed in Section 4.6 can be found in Table D.4. In this context, we compare our proposed *TomoGraphView* framework across various graph topologies and edge connection weighting schemes. For graph topology, we differentiate between local and complete, where local represents a pure spherical structure, while complete represents the spherical structure with additional cross-connections between all nodes in the graph. To this end, we evaluate different edge weighting schemes, including uniform, linear decay, inverse, and inverse-square, which indicate decaying cross-node weights with increasing hop distances.

Appendix E. Slice-wise Feature Aggregation Methods

Detailed quantitative results of our slice-wise feature aggregation benchmark experiment performed in Section 4.7 can be found in Table E.5. Therein, we compare different slice-wise feature aggregation methods, including long short-term memory (LSTM) and medical slice transformer (MST), with our *TomoGraphView* framework across various volume slicing strategies and numbers of views. Numbers are indicated as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds.

Appendix F. 3D Backbones Results

Detailed quantitative results of our comparative large-scale 3D pretrained model analysis performed in Section 4.8 can be found in Table F.6 (frozen backbone performance) and Table F.7 (finetuning performance).

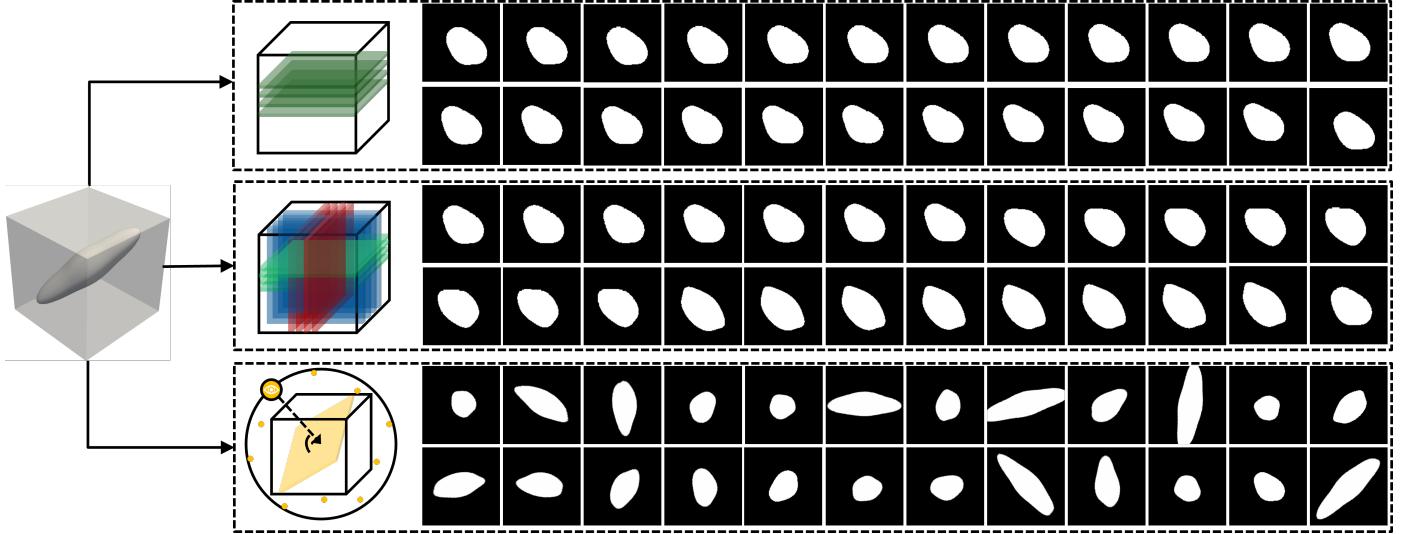


Figure A.9: A visual comparison of the three different volume slicing techniques and their resulting cross-sectional slices for a synthetically generated input volume, showing a target structure of interest which is not aligned with any canonical axis (i.e., axial, coronal, sagittal). The plot shows 2D-axial+ slicing (top), 2.5D+ slicing (middle), and our proposed omnidirectional slicing (bottom), for 24 views each.

Table D.4: Quantitative results for the graph topology analysis comparing different graph topologies and edge weighting schemes. Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds. The best-performing method is indicated in bold, while the second-best is underlined. The column 'Mean' averages the performance across all datasets.

Method	Topology	Weighting	Breast	Brain	Head–Neck	Kidney	Liver	Soft-Tissue	Mean
TomoGraphView	local	uniform	0.6997	0.9667	0.7770	0.7526	0.9666	0.8204	0.8305
	complete	uniform	0.6939	0.9668	0.7480	0.7672	0.9582	0.8475	0.8303
	complete	linear decay	0.6857	0.9653	0.7503	0.7474	0.9614	0.8371	0.8245
	complete	inverse	<u>0.7084</u>	0.9655	0.7527	0.7893	0.9634	<u>0.8441</u>	0.8372
	complete	inverse-square	0.7111	0.9621	<u>0.7549</u>	<u>0.7878</u>	<u>0.9646</u>	0.8249	0.8342

Table E.5: Quantitative results for the slice-wise feature aggregation benchmark comparing different methods, such as long short-term memory (LSTM) and medical slice transformer (MST) with TomoGraphView across volume slicing strategies (i.e., 2D-axial+ and *omnidirectional*) and number of views (i.e., 8, 16, and 24). Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds. The best-performing method is indicated in bold, while the second-best is underlined. The column 'Mean' averages the performance across all datasets.

Method	Volume Slicing	Views	Breast	Brain	Head–Neck	Kidney	Liver	Soft-Tissue	Mean
LSTM	2D-axial+	8	0.6437	0.8826	0.6048	0.6669	0.8923	0.7254	0.7305
		16	0.6505	0.8957	0.6207	0.7072	0.7866	0.7374	0.7330
		24	0.6077	0.8922	0.6467	0.7546	0.8275	0.7388	0.7446
LSTM	Omnidirectional	8	0.6980	0.9419	0.6510	0.6705	0.9133	0.7837	0.7764
		16	0.6731	0.9497	0.6670	0.6733	0.9137	0.7696	0.7744
		24	0.6303	0.9580	0.7179	0.6861	0.8976	0.8280	0.7863
MST	2D-axial+	8	0.6497	0.9232	0.5949	0.7124	0.9002	0.7696	0.7583
		16	0.6539	0.9308	0.6294	0.7563	0.9081	0.7824	0.7768
		24	0.6014	0.9346	0.6539	0.7558	0.9206	0.7815	0.7746
MST	Omnidirectional	8	0.7037	0.9655	0.6443	0.7573	0.9497	0.8295	0.8083
		16	<u>0.7164</u>	0.9663	0.7225	0.7286	0.9238	0.8452	0.8171
		24	0.7029	<u>0.9669</u>	0.7050	0.7751	0.9254	0.8434	0.8198
TomoGraphView	Omnidirectional	8	0.7336	0.9678	0.6780	<u>0.7878</u>	<u>0.9600</u>	0.8399	0.8279
		16	0.6907	0.9666	<u>0.7244</u>	<u>0.7479</u>	<u>0.9585</u>	0.8389	<u>0.8212</u>
		24	0.7084	0.9655	0.7527	0.7893	0.9634	0.8441	0.8372

Table F.6: Quantitative results for the frozen backbone benchmark comparing different methods relying on different pretraining schemes, such as Foundation Model for Cancer Imaging Biomarkers (FMCIB), Models Genesis, SwinUNETR, Versatile Imaging SegmenTation and Annotation model (VISTA3D), Volume Contrast (VoCo) with *TomoGraphView*. Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC), balanced accuracy (ACC), F1-score, and Matthews correlation coefficient (MCC), across 5 folds. The best-performing method per dataset is indicated in bold, while the second-best is underlined. The last section, indicated as ‘Mean’, averages the performance across all datasets and performance metrics.

Dataset	Modality	Method	AUROC	ACC	F1-Score	MCC
Breast Tumors	MRI	FMCIB	0.6197	0.5427	0.5057	<u>0.1016</u>
		ModelsGenesis	0.4527	0.5017	0.3537	0.0024
		SwinUNETR	<u>0.6207</u>	<u>0.5445</u>	<u>0.5303</u>	0.0876
		VISTA3D	0.4349	0.4592	0.3950	-0.0844
		VoCo	0.4957	0.5149	0.5049	0.0321
		TomoGraphView (ours)	0.7214	0.6573	0.6553	0.3177
Brain Tumors	MRI	FMCIB	<u>0.8985</u>	<u>0.7921</u>	<u>0.8587</u>	<u>0.5716</u>
		ModelsGenesis	0.8708	0.6257	0.7895	0.3037
		SwinUNETR	0.8583	0.7694	0.7979	0.4646
		VISTA3D	0.6741	0.5505	0.5843	0.0795
		VoCo	0.7267	0.6719	0.7962	0.3569
		TomoGraphView (ours)	0.9639	0.9131	0.9066	0.7443
Head–Neck Tumors	CT	FMCIB	0.5733	0.5407	0.5448	0.1027
		ModelsGenesis	0.5645	0.5040	0.2805	0.0240
		SwinUNETR	<u>0.5910</u>	0.5609	0.5532	0.1424
		VISTA3D	0.5686	0.5258	0.4875	0.0452
		VoCo	0.5719	<u>0.5759</u>	<u>0.6035</u>	0.1499
		TomoGraphView (ours)	0.6999	0.6464	0.6711	0.2924
Kidney	CT	FMCIB	0.6828	0.6392	0.6785	0.2706
		ModelsGenesis	0.6985	0.5667	0.5181	0.1279
		SwinUNETR	<u>0.7466</u>	<u>0.7112</u>	0.7520	<u>0.4282</u>
		VISTA3D	0.6888	0.6102	0.5508	0.2614
		VoCo	0.6801	0.6535	0.6973	0.3111
		TomoGraphView (ours)	0.7750	0.7245	<u>0.7291</u>	0.4412
Liver Tumors	CT	FMCIB	<u>0.8711</u>	<u>0.8082</u>	<u>0.8060</u>	<u>0.6208</u>
		ModelsGenesis	0.6621	0.5695	0.5013	0.1743
		SwinUNETR	0.7249	0.6696	0.6641	0.3406
		VISTA3D	0.6221	0.5962	0.5941	0.2040
		VoCo	0.6053	0.5896	0.5812	0.1802
		TomoGraphView (ours)	0.9664	0.8956	0.8959	0.7939
Soft-Tissue Tumors	MRI	FMCIB	0.6567	0.5882	0.6499	0.2010
		ModelsGenesis	0.5708	0.5432	0.5308	0.0928
		SwinUNETR	<u>0.6926</u>	<u>0.5595</u>	0.6841	0.1583
		VISTA3D	0.6013	0.5380	0.6495	0.0897
		VoCo	0.6651	<u>0.6327</u>	<u>0.7151</u>	<u>0.2737</u>
		TomoGraphView (ours)	0.8426	0.7340	0.7957	0.4815
Mean	—	FMCIB	0.7170	0.6518	0.6739	0.3114
		ModelsGenesis	0.6366	0.5518	0.4956	0.1209
		SwinUNETR	0.7057	0.6358	0.6636	0.2703
		VISTA3D	0.5983	0.5466	0.5435	0.0992
		VoCo	0.6241	0.6064	0.6497	0.2173
		TomoGraphView (ours)	0.8282	0.7618	0.7756	0.5118

Table F.7: Quantitative results for the finetuning performance of the best-performing frozen backbone, Foundation Model for Cancer Imaging Biomarkers (FMCIB). We employ different warmup durations, along with various batch sizes (BS), and compare their performance with our *TomoGraphView* framework. Results are represented as the mean Area under the Receiver Operating Characteristic Curve (AUROC) value across 5 folds. The best-performing method is indicated in bold, while the second-best is underlined. The column 'Mean' averages the performance across all datasets.

Model	Warmup	BS	Params	Brain	Soft-Tissue	Breast	Head–Neck	Kidney	Liver	Mean
FMCIB (frozen)	n/a	16	100k	0.8985	0.6567	<u>0.6197</u>	0.5733	<u>0.6828</u>	0.8711	<u>0.7170</u>
FMCIB (finetune)	20	256	184M	<u>0.9147</u>	<u>0.7322</u>	0.5049	0.6286	0.5730	0.8708	0.7040
	50	256	184M	0.8981	0.7230	0.5050	0.6308	0.6162	0.8636	0.7061
	100	256	184M	0.9049	0.7013	0.5193	<u>0.6403</u>	0.6589	0.8639	0.7147
	100	16	184M	0.9016	0.6615	0.5540	0.6386	0.6231	<u>0.8811</u>	0.7100
	TomoGraphView (ours)	n/a	16	100k	0.9626	0.8485	0.7491	0.6873	0.7887	0.9603