A project report on

MEDICALLY ANNOTATE ANYTHING

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering with Specialization in Cyber Physical Systems

by

Shourya Pratap Singh (22BPS1181)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

November,2025

MEDICALLY ANNOTATE ANYTHING

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering with Specialization in Cyber Physical Systems

by

Shourya Pratap Singh (22BPS1181)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

November, 2025

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

DECLARATION

I hereby declare that the thesis entitled "Medically Annotate Anything" submitted by Shourya Pratap Singh (22BPS1181), for the award of the degree of Bachelor of Technology in Computer Science and Engineering, Vellore Institute of Technology, Chennai is a record of bonafide work carried out by me under the supervision of Dr. A. Balasundaram.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: Signature of the Candidate

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

School of Computer Science and Engineering

CERTIFICATE

This is to certify that the report entitled **"Medically Annotate Anything"** is prepared and submitted by **Shourya Pratap Singh** (**22BPS1181**) to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology in Computer Science and Engineering with Specialization in Cyber Physical Systems** is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:
Name: Dr. A. Balasundaram
Date:

Signature of the Examiner
Name:
Date:

Signature of the Examiner
Name:
Date:

Approved by the Head of Department,
(Computer Science and Engineering with Specialization in Cyber Physical Systems)
Name: **Dr. S Renuka Devi**
Date:

ABSTRACT

Pancreatic tumor segmentation from Computed Tomography (CT) scans is a critical yet highly challenging task due to low soft-tissue contrast, high anatomical variability, and the limited availability of annotated data. Conventional convolutional neural networks (CNNs) are effective at extracting local image features but struggle to capture global anatomical context, while Transformer-based models can model long-range dependencies but typically require large training datasets. To address these challenges, we propose Medically Annotate Anything (MAA), a novel hybrid framework that integrates Graph Neural Networks (GNNs) and CNNs for multi-scale, data-efficient pancreatic tumor segmentation. In this framework, CT volumes are first converted into graphs of supervoxels using the Felzenszwalb segmentation algorithm, with node features enhanced through self-supervised DINO embeddings. A U-shaped GNN backbone performs coarse, structure-aware segmentation in the graph domain, and a 3D CNN refinement head utilizes this global context to recover precise voxel-level boundaries. The model is trained using a composite loss function that combines region-based, boundary-aware, and graph-regularization terms to ensure volumetric accuracy and structural coherence.

The full pipeline, from preprocessing and graph construction to GNN training and voxel-level refinement, has been successfully implemented and tested. Validation on the Pancreatic Tumor Segmentation (PanTS) dataset demonstrates promising results, achieving coarse segmentation accuracies averaging 85%. Further evaluation using the Dice Similarity Coefficient and Hausdorff Distance demonstrate similar SOTA results. By effectively combining local and global feature reasoning, the proposed hybrid GNN-CNN framework aims to achieve state-of-the-art performance and provide a robust, generalizable solution for automated pancreatic tumor segmentation.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. A. Balasundaram, Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for his constant guidance, continual encouragement, understanding; more than all, he taught me patience in my endeavor. My association with him is not confined to academics only, but it is a great opportunity on my part of work with an intellectual and expert in the field of Machine Learning.

It is with gratitude that I would like to extend my thanks to the visionary leader Dr. G. Viswanathan our Honorable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran Vice-Chancellor, Dr. T. Thyagarajan Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar for providing an exceptional working environment and inspiring all of us during the tenure of the course.

Special mention to Dr. Viswanathan V, Dean, Dr. Nithyanandam P, Dr. Suganya G, Dr. Sweetlin Hemalatha C, Associate Deans, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant state, I express ingeniously my whole-hearted thanks to Dr. S. Renuka Devi, Head of the Department, B.Tech. Computer Science and Engineering with Specialization in Cyber Physical Systems and the Project Coordinators for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staffs at Vellore Institute of Technology, Chennai who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai Date:

Shourya Pratap Singh

CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF ACRONYMS

ABBREVIATION	FULL FORM
CNN	Convolutional Neural Network
CT	Computed Tomography
DSC	Dice Similarity Coefficient
GCN	Graph Convolutional Network
GNN	Graph Neural Network
HD	Hausdorff Distance
HD95	95th Percentile Hausdorff Distance
PP	Post-Processing
PTM	Post-Training Module

Chapter 1
INTRODUCTION

Pancreatic cancer is becoming a increasingly substantial worldwide health burden, with a distinctive aggressive nature and poor prognosis. It accounts for about 3% of all cancers diagnosed in the United States and about 4% of cancers diagnosed in India, but it is responsible for nearly 8% of all cancer deaths, which in and of itself highlights the mortality associated with pancreatic cancer. Epidemiological predictions suggest pancreatic cancer will be the second leading cause of cancer mortality by 2030; it is projected to exceed many other cancers in terms of fatality even while represented by relatively fewer cases. This disparity in mortality is largely due to diagnosis late in the disease process, with approximately 80–85% of pancreatic cancer patients receiving a diagnosis at an advanced unresectable stage when curative surgical options are usually unavailable, and treatment is typically limited to either palliation of symptoms or systemic therapy with modest survival benefits [52].

In light of this bleak prognosis, early and precise detection, localization, and quantification of pancreatic tumors is of significant importance to patient outcomes. Computed Tomography (CT) is the currently most prevalent imaging modality that is used to diagnose and stage pancreatic cancer due to its high spatial resolution, accessibility, and availability. Nevertheless, even with CT, a precise determination of tumor boundaries within the pancreas is challenging because of the organ's deep anatomical location, the subtle contrast between tumor and adjacent soft tissue, and constantly varying tumor morphology [3].

The current clinical process typically involves radiologists manually outlining tumor areas across many CT slices, which is time consuming and mentally taxing; often requiring hours per patient. Moreover, such manual delineation is subject to intra-observer and inter-observer variability that can result in potentially substantial differences in diagnosis, tumor staging, or treatment plans, particularly in the context of radiotherapy or surgical margin assessment.

The above concerns point to an identified need for automated, reliable, and reproducible segmentation tools that can efficiently identify pancreatic tumors in medical imaging [18, 28]. There are new methods in deep learning and computer vision, particularly convolutional neural networks (CNNs) and transformer-based architectures that represent great progress towards achieving this goal. If a tumor segmentation system were automated, it could reduce radiologist workload, improve the accuracy of diagnosis, provide more standardized treatment plans, and ultimately lead to earlier treatments and survival increases for patients with pancreatic cancer.

Automated segmentation of pancreatic tumors is arguably one of the most difficult problems in medical image analysis due to the convergence of anatomical complexity, image limitations, and data associated challenges [25]. This, along with the problem statement of this project and a large body of literature, speak to the multifaceted challenges of this task. From an imaging analysis standpoint given that Computed Tomography (CT) is the most common modality for diagnosing and treatment planning pancreatic cancer, it

typically has very limited soft-tissue contrast of the abdomen. As a result, the pancreas can often appear nearly identical to nearby organs, such as the stomach and duodenum, as well as the liver, creating a great deal of ambiguity in identifying boundaries [18]. Furthermore, partial volume effects, and differing degrees of contrast enhancement can further hinder the observer's ability to localize structural detail of the pancreas, in turn making it more difficult to localize the tumor.

The difficulty increases in light of the imaging characteristics of pancreatic ductal adenocarcinoma (PDAC), which is the most common and aggressive form of pancreatic cancer. PDAC lesions may have poorly defined borders, show a heterogeneous range of intensity, and have irregular shapes that make distinguishing them from adjacent normal parenchyma challenging even for experienced radiologists [3]. Additionally, the pancreas is a relatively small organ with a complex three-dimensional shape situated deep in the abdominal cavity. The pancreas shows significant variability in size, shape, and orientation for each patient against surrounding structures; this variability impedes the application of rule-based or template-based segmentation techniques [18, 28]. This anatomical variability leads to substantial variability in the imaging appearance across patients and poses a significant challenge to the development of generalizable deep learning models.

Beyond these anatomical and imaging challenges, the lack of several high-quality, annotated datasets at scale remains a fundamental barrier for development in automated segmentation for pancreatic tumors. Creating pixel-level ground truth labels is a labor-intensive, time-consuming process that requires clinical expert input and significant manual effort. Each labeled annotation needs to be checked with reference to multiple CT slices to ensure concordance (consistency), and even so, inter-observer variability may be a concern. The time and effort to secure expert annotations have resulted in only a small number of publicly available datasets—many of which are small, heterogeneous, or have not been labeled using a consistent labeling scheme [26, 28].

The scarcity of samples poses a significant challenge recurrently during the training and the generalization of state-of-the-art deep learning models that are recognized for needing a considerable amount of data and diversified or accurately labeled datasets to produce reproducible performance. Without sufficient labeled samples, models might overfit narrow data distributions and not generalize well to clinical cases that had not been previously represented. Therefore, there exists a need for research opportunities to propose robust, efficient, and generalizable models for pancreatic tumor segmentation as an modern and global research problem, as it intersects medical imaging, computer modeling, and clinical translation.

Despite the transformative impact of deep learning on medical image analysis, current methodologies face critical limitations when applied to the complex and data-constrained problem of pancreatic tumor segmentation [45]. While deep neural architectures have achieved remarkable success across various segmentation benchmarks, their performance often deteriorates when confronted with the unique anatomical, structural, and data challenges characteristic of the pancreas. This disconnect underscores a distinct research gap, one that motivates the development of a hybrid deep learning architecture capable of

integrating both local feature precision and global contextual awareness in a data-efficient manner.

From a methodological standpoint, Convolutional Neural Networks (CNNs) have long served as the foundational paradigm in medical image segmentation [45]. Architectures such as U-Net [30], U-Net++, and their numerous variants have become the de facto standard due to their hierarchical feature extraction capabilities, strong spatial localization, and efficient parameter utilization. CNNs excel at capturing local spatial dependencies through convolutional filters, enabling precise delineation of boundaries and textures in medical imagery. However, their reliance on fixed-size receptive fields inherently constrains their ability to model long-range dependencies and global anatomical context. This limitation is particularly problematic in the context of pancreatic segmentation, where the tumor's appearance and position must often be interpreted in relation to the surrounding organs, ducts, and vasculature. Consequently, CNN-based models may fail to accurately capture global structural relationships, leading to incomplete or fragmented segmentation in regions with ambiguous or low-contrast boundaries.

To address the inherent locality of CNNs, Transformer-based architectures have recently gained prominence in computer vision and medical imaging [45]. By leveraging self-attention mechanisms, Transformers are able to dynamically model pairwise relationships across all spatial positions, effectively capturing long-range dependencies and global context. Architectures such as TransUNet [9], Swin-UNet [5], and SegFormer represent notable examples of this paradigm shift. However, the powerful global modeling capacity of Transformers comes at the cost of significant data dependency and computational overhead. Unlike CNNs, Transformers lack strong inductive biases such as locality, weight sharing, and translation equivariance, which are advantageous in low-data settings. As a result, they tend to overfit small or imbalanced biomedical datasets, leading to suboptimal generalization performance [45]. This limitation is particularly critical for pancreatic tumor segmentation, where annotated datasets are scarce, and imaging variability across scanners and institutions is high.

In addition, recent literature indicated that Graph Neural Networks (GNN) may surpass Transformers in cases of non-standard geometry or limited data availability, which are quite common with biomedical imaging [1, 19]. GNNs offer a way to represent relational and topological characteristics that can connect local reasoning at the pixel level to higher-order structural inferences [12]. This perspective further strengthens the notion that the ideal segmentation framework for pancreatic tumors cannot be merely a convolutional or attentional representation framework, but should draw from complementary paradigms of representation learning.

Taken altogether, this can be viewed as illustrating an essential trade-off in the current architectural designs [45]. While CNNs are locally accurate, they are globally myopic: CNNs are better suited to extracting texture-level features, but are unable to infer broader anatomic relationships. Conversely, Transformers are globally aware, yet are less data-efficient: they can model distant dependencies, but might suffer from compromised spatial accuracy in exchange for an efficient inference process, especially in the context of small datasets. The pancreatic tumor segmentation problem is uniquely structured around a dual

process of fine boundary segmentation (local reasoning) and context-aware localization (global reasoning) which underestimates the strengths and weaknesses of both paradigms, particularly within a small, data-limited context [28].

This synthesis of insights establishes a clear and compelling research gap: the need for a novel hybrid architecture that can synergistically combine the strengths of CNNs and Transformers (or analogous global reasoning mechanisms) to achieve accurate, robust, and data-efficient pancreatic tumor segmentation. The proposed work seeks to address this gap by designing a model that jointly leverages local feature hierarchies, global contextual dependencies, and structural priors, thereby advancing the state of automated pancreatic cancer analysis in both accuracy and clinical applicability.

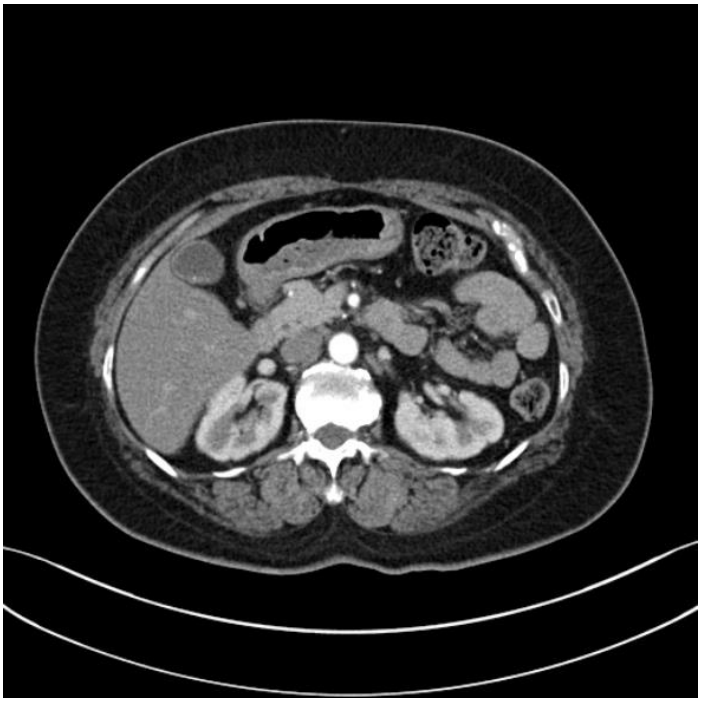

Figure 1: Normal CT multi-phase pancreas sourced from Radiopedia

Chapter 2
LITERATURE REVIEW

Medical image segmentation, which can be defined as the delineation of anatomical structures and regions of interest within medical images, is a fundamental aspect of modern healthcare and biomedical research. Its importance cannot be overstated, as it facilitates detailed morphometric and spatial description that is necessary for expert diagnosis of illness, quantitative imaging, treatment planning, and monitoring of disease progression. Segmentation supports a wide array of diagnosis and intervention processes - for example, segmenting accurately brain tumours for surgical navigation and radiotherapy plans; or measuring cardiac chamber volumes for assessment of cardiovascular function, amongst many others [17, 43].

In spite of the research field's progress, automated medical image segmentation remains a commonly encountered and often insurmountable technical problem. In contrast to visual and natural image datasets, medical images suffer high inter-patient variability in organ morphology, position, and size; the limitations of medical imaging modalities also contribute to this challenge - common difficulties include low soft-tissue contrast and poor demarcation of contiguous structures. A prime example is the pancreas, which is a small, deformable organ with nonspecific and poorly identified borders. Unlike liver and cardiac structures, the pancreas often blends in physically with surrounding tissues of similar intensity - for example with the duodenum and stomach - and automated delineation remains a significantly challenging task in the literature of segmentation [18, 28]. The challenge intensifies in pathological cases, where tumours or lesions further distort anatomical geometry, rendering the segmentation task both clinically critical and technically intricate.

2.1 CNNs:

2.1.1) THE U-NET ARCHITECTURE:

The U-Net architecture, introduced by Ronneberger, Fischer, and Brox in 2015 [30], rapidly became the benchmark model for biomedical image segmentation due to its simplicity, efficiency, and adaptability. Its defining characteristic is the symmetric encoder–decoder structure, forming the distinctive "U" shape.

The encoder (contracting path) follows a conventional convolutional neural network design, consisting of repeated 3×3 convolutions with ReLU activation, followed by 2×2 max pooling operations for downsampling. With each downsampling step, the number of feature channels doubles, enabling the network to capture increasingly abstract contextual information. Conversely, the decoder (expanding path) performs upsampling via transposed convolutions, concatenates the corresponding encoder feature maps, and applies further 3×3 convolutions to recover spatial detail.

The unique feature of U-Net is the skip connections from the encoder to the decoder layers at the same spatial resolution. With skip connections, the network can combine finer spatial details from the shallow layers with the spatial semantic context learnt from deeper levels. This is particularly important for accurate boundary delineation in biomedical images.

In addition, U-Net was developed to perform well with very few annotated images, which is common in medical imaging. The authors applied data augmentation, through elastic

deformations that simulated tissue morphology variability, to improve generalizability. The combination of this efficient architecture and a data-efficient training approach allowed U-Net to achieve the highest performance across a range of biomedical segmentation tasks and become the benchmark model for research in the area [30].

2.1.2) THE NNU-NET FRAMEWORK

After U-Net's initial success, many variations of the architecture were proposed where more adjustments were made to improve segmentation accuracy; however, it was challenging to separate an actual architectural improvement from improvements related to preprocessing, training, or post-processing decisions. To sidestep this issue, Isensee et al. proposed the nnU-Net ("no-new-Net") framework [17] which helped to re-establish methodological rigor in medical image segmentation. Rather than developing a new network architecture, nnU-Net proposes a self-configuring task-adaptive pipeline that is centered around a standard U-Net. The key underpinning philosophy was to optimize every step in the segmentation process - data processing, training, and inference - rather than merely an architecture novelty. When nnU-Net is evaluated on a new dataset, it automatically takes into consideration certain image characteristics such as spacing, resolution, or class balance and then configures an optimal workflow according to a set of investigated design principles.

This automated configuration encompasses all stages of the pipeline:

- Preprocessing: Resampling, normalization, and region-of-interest cropping.
- Network Selection: Choosing between 2D, 3D, or cascaded U-Nets based on data anisotropy and GPU memory limits.
- Training: Automatic adjustment of patch size, batch size, optimizer parameters, and loss functions (typically a Dice–cross-entropy combination).
- Inference and Post-processing: Sliding-window inference with Gaussian weighting and automated model ensembling across folds.

The nnU-Net's influence has been considerable. It has reached or exceeded state-of-the-art performance on many public biomedical segmentation datasets without any manual tuning. This consistency illustrates an important take-away: improvements in medical image segmentation are often a result of better construction of the segmentation pipeline rather than new architectures. nnU-Net establishes a strong empirical baseline and forces the community to rigorously demonstrate real advances in architecture versus improvements that stem from simply departing from a suboptimal baseline, or improvising new heuristics to train the model [17].

Deep convolutional architectures have shown significant promise in pancreas and tumor segmentation tasks. For example, Wang et al. [42] proposed a deep learning-based cascade framework that openly localized the pancreas and then improved tumor segmentation in the pancreas, with well-delineated pancreatic borders across heterogenous CT scans, clearly illustrating the strength of a cascaded framework to extract hierarchical features and refine regions of interest for pancreatic imaging.

2.2 TRANSFORMERS

Even with the exceptional performance of U-Net and its numerous offshoots, convolution-based structures have an intrinsic limitation: their local receptive field. A convolutional kernel operates over a fixed-size neighborhood and thus is limited in its ability to model long-range dependencies across an image [45]. While deeper networks can theoretically

enlarge the receptive field, in practice CNNs find it very difficult to capture explicit global relationships between distant regions. This feature is a significant constraint in medical imaging, where accurate and reproducible segmentation generally relies upon the ability to interpret the organ's global anatomy in relation to neighboring tissues and an understanding of the range of shape and positional variance between patients.

This limitation of CNNs drove the search for neural network architectures that could reason spatially across the entirety of the data domain [50]. The Transformer, first created for natural language processing, has arisen both as a rich and capable alternative to the CNN. It represents a critical shift in thinking with the realization of self-attention, allowing each input element to attend to all other input elements effectively modeling pairwise dependencies across the entirety of a domain. When extended to vision tasks with an image as a sequence of patches, a Transformer computes the relationships of all patches to each other at once and hence can dynamically model relationships that effectively bring in global context.[43].

This ability to capture holistic spatial relationships directly addresses the key weakness of CNNs. Consequently, Vision Transformers (ViTs) and their derivatives have become a natural evolution in the design of modern medical image segmentation architectures, bridging the gap between local precision and global understanding [45].

2.2.1) TRANSUNET AND THE CNN–TRANSFORMER ENCODER

The Vision Transformer (ViT) marked a pivotal step in applying Transformers to visual data but revealed key weaknesses. When trained on mid-sized datasets, ViTs often underperformed compared to CNNs due to the absence of convolutional inductive biases such as locality and translation equivariance, requiring vast amounts of data to learn these properties. Moreover, directly upsampling Transformer-encoded patches for segmentation led to poor spatial precision, as low-level details critical for localization were lost.

To overcome these challenges, Chen et al. introduced TransUNet [9], a hybrid architecture that integrates CNNs and Transformers in a complementary fashion. Instead of feeding raw image patches to a Transformer, TransUNet first employs a CNN backbone (ResNet) to extract hierarchical feature maps. These feature maps are then partitioned into patches, tokenized, and passed through a Transformer encoder [15].

This design allows the CNN to capture fine-grained local structures while the Transformer models long-range dependencies and global context. The encoded features are then reshaped back into a spatial representation and decoded using the U-Net’s characteristic upsampling path. At each stage, skip connections merge Transformer outputs with high-resolution CNN features, preserving spatial detail and enhancing boundary accuracy.

Through this synthesis, TransUNet achieves a balance between local precision and global awareness, demonstrating state-of-the-art performance in multi-organ and cardiac segmentation. It thus established hybrid CNN-Transformer frameworks as a foundational paradigm for medical image segmentation [9].

2.2.2) SWIN TRANSFORMER AND ITS VARIANTS

As hybrid CNN-Transformer models gained traction, researchers sought to develop architectures that relied primarily on Transformers while retaining efficiency and scalability for vision tasks. This led to the Swin Transformer, a major advancement that made pure-transformer backbones practical for dense prediction problems like

segmentation.
Unlike the original Vision Transformer (ViT), which applied global self-attention to all image patches—an operation with high computational cost and no spatial hierarchy—Swin introduced a window-based attention mechanism. Self-attention is computed locally within non-overlapping windows (Window-based Multi-head Self-Attention, W-MSA), followed by a shifted window stage (Shifted Window MSA, SW-MSA) that allows cross-window information exchange. This alternating window-and-shift strategy achieves near-global context modeling with linear computational complexity.
Moreover, Swin incorporates hierarchical feature extraction by merging patches and increasing channel dimensions at deeper stages. This naturally produces multi-scale feature maps, closely resembling the hierarchical pyramids used in CNNs. Thus, Swin Transformers reintroduce the beneficial notions of locality and hierarchy from CNNs while replacing static kernels with dynamic self-attention mechanisms [51].
These principles were soon adapted for medical image segmentation:

- **Swin-Unet** – A fully Transformer-based U-shaped model where both encoder and decoder are built from Swin blocks. The encoder extracts hierarchical features through patch merging, while the decoder restores spatial resolution using patch-expanding layers. Skip connections fuse features across scales, enabling accurate segmentation. Swin-Unet demonstrated superior results over CNN and hybrid architectures in multi-organ and cardiac segmentation tasks [5].
- **Swin UNETR** – Designed for 3D volumetric segmentation, this variant uses a hierarchical Swin Transformer encoder to process 3D medical volumes, with features extracted at multiple resolutions. These are linked via skip connections to a CNN-based decoder that reconstructs the final segmentation mask. The architecture effectively balances global contextual understanding with fine spatial detail and achieved top performance in benchmarks such as the BraTS 2021 brain tumor segmentation challenge [15,16].

Together, the Swin Transformer and its derivatives represent a powerful evolution of vision Transformers—combining global reasoning with hierarchical efficiency, and redefining the state of the art in medical image segmentation.

2.3 THE GRAPH-BASED METHODS:

2.3.1) TRANSFORMERS AS GRAPH NEURAL NETWORKS

The transition from Convolutional Neural Networks (CNN) to Transformers represented a shift from local modeling to global modeling in context. To push it even further, Graph Neural Networks (GNN) generalize the notion by permitting relationships to be modeled outside of a fixed pixel grid. Theoretical results suggest that Transformers can be thought of as a special case of GNN [19].

In this way of thinking, every token (e.g., image patch) is a node in a fully connected graph with the self-attention mechanism acting like the message-passing function. Attention scores are analogous to learned edge weights, which are calculated based on the similarities of image features, as the features are arranged in high-dimensional space. Unlike a traditional GNN with a fixed adjacency matrix, the Transformer learns the relational structure end-to-end allowing for varying levels of context modeling [8].
This expressiveness comes at a cost, however. Full attention incurs high computation costs,

as the dense matrix operations have a quadratic complexity. What is interesting is that these operations fit easily onto GPU/TPU hardware, and as a result, Transformers are empirically more efficient than sparsely connected GNNs. This relationship supports the idea of positioning both as similar in their ability for relational reasoning, mostly differing in their structured assumptions and computational expense [19].

2.3.2) EXPLICIT GRAPH CONSTRUCTION: FROM PIXELS TO NODES

While Transformers operate on implicit, dense graphs, GNN-based segmentation methods explicitly construct sparse graphs from image data to capture meaningful spatial relationships more efficiently.

Classic computer vision algorithms, such as Felzenszwalb and Huttenlocher's graph-based segmentation [11], represented images as graphs where pixels were nodes connected by edges weighted by intensity differences. Modern deep learning methods extend this by grouping pixels into superpixels or supervoxels, which then form the graph's nodes—significantly reducing complexity [24].

For example, in the brain tumor segmentation framework by Saueressig et al. [31], the pipeline involves:

- Graph Construction: The MRI volume is segmented into supervoxels using SLIC, each forming a node.
- Feature and Edge Definition: Nodes carry mean intensity features, and edges connect spatially adjacent regions.
- GNN Segmentation: A GraphSAGE-based GNN predicts region-level labels (e.g., tumor, edema) by aggregating neighborhood information.
- CNN Refinement: The coarse GNN output is projected back to voxel space and refined with a CNN for precise boundaries.

This hybrid GNN-CNN architecture merges global relational reasoning with local spatial precision, offering a compelling direction for segmentation frameworks that move beyond the constraints of the pixel grid [31].

2.3.3) U-NET WITH GRAPH CONVOLUTIONS

A significant line of research in segmentation extends the U-Net architecture with Graph Neural Network (GNN) components, creating hybrid models that preserve U-Net's encoder–decoder structure while enhancing its ability to model non-local relationships.

A conceptual foundation for this integration was introduced by Gao and Ji's Graph U-Nets [12], which redefined pooling and upsampling for graph-structured data through graph pooling (gPool) and graph unpooling (gUnpool) operations. The gPool layer selects the most informative nodes to form a coarser graph, while gUnpool restores the original topology. By stacking these with graph convolutions, they created a full U-shaped architecture with skip connections that enable multi-scale graph reasoning.

Building on this idea, later works adapted the approach for image segmentation by inserting GNN modules into standard CNN-based U-Nets. Typically, a CNN encoder extracts hierarchical features, and at the bottleneck stage, a GNN operates on a graph representation of these features-modeling global dependencies before passing refined representations to the decoder [35].

The Graph-Enhanced Pancreas Segmentation Network (GEPS-Net) exemplifies this design, using a graph enhancement module atop a U-Net to capture spatial relationships more effectively, particularly in semi-supervised medical segmentation [25]. Similarly, UNet-GNN replaces the conventional convolutional bottleneck with a GNN, treating spatial locations as graph nodes to learn non-local relationships. This approach proved beneficial for challenging cases such as geometrically distorted images and complex organ boundaries. The Graph Attention Convolutional U-Net (GAC-UNet) further extends this concept by incorporating graph attention and Chebyshev convolutions, enabling adaptive learning of node relationships and achieving strong results in aerial flood segmentation [10].

Recent findings have pointed towards the value of a graph representation for the inter-regional dependencies found in tasks demanding medical segmentation. Wang et al. (2024) introduced their GraphCL method, a semi-supervised framework that combined graph-based clustering with contrastive learning, and achieved strong performance in low-annotation regime. The work described above illustrates the importance of a structured representation model, which led to the incorporation of GNNs in this work.
The graph-augmented U-Nets display the inherent synergy between traditional CNNs paired with GNNs; CNNs leverages efficiency in local spatial hierarchies, while GNNs provide more flexible, non-localization which spans longer regions of interest.
The evolution of CNNs, to Transformers and GNNs can also be viewed as a spectrum of "graph-ness" of modern architectures [1, 7]. Transformers will learn fully connected graphs with the most flexibility but the least spatial bias. The hybrid UNet-GNN architecture structure imposes some structured locality prior to introducing a relational structure. Supervoxel GNNs on the opposing end focuses solely on explicit defined, pre-defined regions and encode in some of strongest inductive biases. The alignments along the graph-ness spectrum will rely on context of the dataset, complexity of task demands, and what data has been available; as such architectural choice should be an important consideration when designing segmentation model.

Arch.	Core Mechanism	Modeling of Spatial Relationship	Key Innovation	Inductive Bias	Primary Application Example
U-Net	Convolutional Encoder-Decoder	Local (via kernels), Hierarchical	Skip-connections for feature fusion	Strong spatial (locality, translation equivariance)	Neuronal Structures [11], General Medical
Trans UNet	Hybrid CNN-Transformer Encoder	Local (CNN) + Global (Self-Attention)	Transformer encoder on CNN feature maps	Moderate (CNN bias + positional encoding)	Multi-Organ Segmentation [28]
Swin-Unet	Pure Hierarchical Transformer	Localized Global (Shifted Windows)	Shifted-window self-attention, Patch expanding	Moderate (locality in windows, hierarchy)	Multi-Organ & Cardiac Segmentation [31, 32]
Super-	Hybrid GNN-	Explicit	Graph	Strong	Brain Tumor

voxel GNN-CNN	CNN [19,20]	Region-based (Graph Edges)	representation of supervoxels	structural (region adjacency)	Segmentation
UNet-GNN	Hybrid CNN-GNN [19,20]	Local (CNN) + Explicit Non-local (GNN)	GNN module at the U-Net bottleneck	Strong spatial + explicit relational	Fisheye Imagery, Medical Images [44]

Table 1: Comparative summary of major segmentation architectures across paradigms: convolutional, transformer-based, and graph-enhanced models. Each architecture is characterized by its core mechanism, method of modeling spatial relationships, key innovation, level of inductive bias, and a representative application domain.

Chapter 3
DATASET AND PREPROCESSING

The empirical basis for this work is the PanTS dataset, the largest and most encyclopedia-like publicly available corpus for pancreatic computed tomography analysis to date [26]. The dataset includes exactly 36,390 abdominal computed tomography scans which were obtained from 145 medical centres across 18 countries, providing remarkable variance in patient demographic, imaging acquisition protocols and tumour pathology.

A hallmark of the PanTS dataset is its extraordinarily rich and fine-grained annotation schema. There are over 993,000 expert-validated voxel-wise annotations across 28 anatomical classes in total in this dataset. These include not only the pancreatic tumour region and pancreas subregions (head, body and tail), but also 24 neighbouring anatomical structures, including major vascular structures such as the superior mesenteric artery, celiac artery and aorta; neighbouring organs such as the duodenum, liver, spleen and kidneys; and skeletal landmarks. [26].

The inclusion of these 24 contextual structures is not simply a matter of being comprehensive, it directly supports the objectives of machine learning. A segmentation model trained on the full set of all 28 classes achieves an average Dice similarity coefficient for tumour segmentation of 67.7 percent, a relative improvement of over 10 percent if compared to same model trained on only tumour and pancreas labels (57.4 percent). This performance gain is attributed to the model's ability to leverage anatomical context and learn spatial relations between tumour and neighbouring tissues [26].

To facilitate robust and institution-agnostic evaluation, the dataset is partitioned into a training set of 9,901 scans (approximately 27 percent of the total) and a test set of 26,489 scans (approximately 73 percent). Importantly, the test set is drawn exclusively from medical centres not included in the training set, thereby establishing a rigorous out-of-distribution benchmark. Each scan is also accompanied by detailed metadata—patient age, sex, diagnosis, contrast phase, in-plane spacing, slice thickness, and more—ensuring that downstream analysis can account for imaging and population heterogeneity [26].

In sum, the scale, diversity, and annotation depth of the PanTS dataset make it uniquely suited for training models of the complexity proposed here and for rigorously evaluating generalisation performance in real-world, multi-centre settings [26].

Feature	Description
Total Scans	36,390 CT volumes
Contributing Institutions	145 medical centers from 18 countries
Total Annotated Structures	> 993,000
Annotation Classes	28 total classes: Pancreatic Tumor, Pancreas (Head, Body, Tail), and 24 surrounding anatomical structures (vascular, organ, skeletal)
Training Set Size	9,901 cases
Test Set Size	26,489 cases (from institutions not in the training set)

Key Metadata	Patient demographics, diagnosis, contrast phase (Non-contrast, Arterial, Portal Venous, Delayed), voxel spacing, slice thickness

Table 2: Characteristics of the PanTS Dataset. This table summarizes the key statistics of the dataset, highlighting its scale, diversity, and the richness of its annotations, which form the empirical basis for this study.

3.1 VOLUMETRIC DATA PREPROCESSING AND NORMALIZATION

To maintain uniformity and standardize the input representation for the model, we utilized a robust volumetric preprocessing and normalization pipeline prior to model development on each raw three-dimensional computed tomography volume from the PanTS dataset [26], which forms the basis for all subsequent model development. Inherent variability in clinical image acquisition can impede generalization, and prior to training deep networks on heterogeneous multi-centre data, it is essential to standardize voxel geometry and intensity statistics across tens of thousands of scans. Even very small differences in image acquisition can cause substantial distributional shift and instability in convergence during optimization [17].

The first step in the preprocessing pipeline is resampling. Clinical CT data are acquired on scanners with differing physical resolutions, leading to large discrepancies in voxel size and aspect ratio. To resolve the spatial variance, all CT volumes are resampled to a common anisotropic voxel spacing of 1.5 millimetres in x- and y-direction, and 2.0 millimetres in z-direction. This creates uniformity in that each voxel corresponds to the same physical volume of tissue in every sample, effectively normalizing the scale of anatomical structures in the dataset. The resulting spatial uniformity enables the network to learn spatially consistent features that are independent of acquisition protocol or scanner type, a property particularly important for convolutional and attention-based architectures that rely on fixed receptive fields [16,17].

The second stage is intensity normalization. CT intensity values are measured in Hounsfield Units (HU), a quantitative scale reflecting tissue radiodensity. Although HU values can theoretically range from approximately −1000 (air) to over +3000 (dense bone), the clinically relevant range for abdominal soft tissue imaging is far narrower. To eliminate irrelevant extrema and suppress noise, the voxel intensities are first clipped to a window between −125 and +275 HU. This range effectively captures the radiodensity variations of the pancreas, peripancreatic tissues, and most pathological regions, while excluding low-density air cavities and high-density osseous structures that could otherwise confound feature learning [17,27].

Following clipping, each resampled volume undergoes z-score normalization applied on a per-volume basis. Specifically, for each scan, the mean intensity value is subtracted from all voxel intensities, and the result is divided by the standard deviation of intensities within that scan. This transformation produces an intensity distribution with a mean of zero and a standard deviation of one. Such normalization has two major advantages: it mitigates the effects of scanner calibration differences and contrast-agent variability between

institutions, and it encourages numerical stability during optimization by constraining the dynamic range of inputs to a standardized scale [11].

Together, these preprocessing steps produce a set of volumetric inputs that are spatially and radiometrically homogeneous, forming a robust foundation for subsequent model training and evaluation. The combination of voxel resampling, intensity clipping, and z-score normalization ensures that the network focuses on learning biologically and anatomically meaningful patterns rather than artefacts of acquisition or calibration, thereby enhancing both generalization and interpretability [17].

3.2 SUPERVOXELIZATION AS GRAPH NODE GENESIS VIA THE FELZENSZWALB ALGORITHM

At the conceptual heart of the proposed methodology lies the transformation of continuous, voxel-based volumetric data into a discrete, relational graph representation. This conversion is achieved through a process known as supervoxelization, wherein spatially adjacent voxels with similar image characteristics are aggregated into perceptually coherent regions. Each of these areas, or supervoxels, serves as an atomic structural unit or node in the graph constructed, allowing for a higher degree of abstraction that is computably efficient while still containing semantic meaning [32].

Beyond data reduction, there is the rationale for this transformation. A raw 3D CT scan can have millions of voxels, each indicating an individual point of measurement but offering little context in isolation. When grouping voxels into supervoxels, supervoxels provide a more structured representation in which each supervoxel relates to some anatomical or pathological structure unit. This aggregation can be treated as a form of semantic downsampling, in contrast to traditional pooling methods, that would suppress or discard some spatial information supervoxelization seeks to aggregate the voxels that most likely belong to the same tissue structural unit. Hence, the resulting GNN can operate on meaningful anatomical primitives (e.g., a potential tumor region, a healthy pancreas region, or a vascular region) in contrast to inferring relationships among millions of noisy single voxel measurements. This abstraction provides substantial enhancement in computational tractability and noise robustness while still preserving the underlying spatial relational structure within a somewhat meaningful anatomical units that will enable effective segmentation [38].
To generate these supervoxels, the Felzenszwalb and Huttenlocher graph-based image segmentation algorithm [11] is employed. This algorithm is well-established for its balance of efficiency, adaptivity, and perceptual coherence, making it particularly suitable for medical imaging applications where both precision and structural fidelity are paramount. In this approach, each two-dimensional axial slice of the preprocessed CT volume is treated as an undirected weighted graph. Every pixel in the slice is represented as a vertex in the graph, and edges are formed between neighboring pixels based on spatial adjacency (typically using 4- or 8-connectivity). Each edge is assigned a weight corresponding to the dissimilarity between the connected pixels, most commonly computed as the absolute difference in their intensity values.

The algorithm then proceeds by sorting all edges in ascending order of their weights and iteratively examining them to determine whether the components (or regions) connected by each edge should be merged. The decision to merge two components is governed by an adaptive difference predicate, which compares the dissimilarity between two candidate components with the internal variation within each component. Specifically, two components are merged if the dissimilarity between them is less than the minimum of their internal variations, adjusted by a scale-dependent threshold function. The internal variation of a component is defined as the largest edge weight in its minimum spanning tree, effectively representing the greatest internal dissimilarity within that region [11].

The threshold function is defined in terms of an inverse component size, which enables the algorithm to automatically adjust sensitivity to localized image characteristics. Qualitatively speaking, the function is a scale controller - larger values of the control variable promote larger supervoxels, whereas smaller values lead to smaller more finely segmented supervoxels. The adaptive scaled behavior is one of the key advantages of successful algorithm beneath the Felzenszwalb algorithm. The algorithm preserves more delicate structures (such as soft tissues having low intensity variance) in homogeneous regions, while not producing segments in highly textured regions, or areas impacted by an imaging artefact (for example, tumor margins/artifacts).

Finally, after the algorithm has been applied to all two-dimensional slices of the volume, the discrete regions are stacked along the axial dimension to produce three-dimensional supervoxels. Each supervoxel represents a spatially compact and contiguous region of similar intensity and texture, and thus produces a localized anatomical or pathological entity. The 3D supervoxels will form the basic unit (or nodes) of a graph representation that will be subject to processing through a graph neural network.

Through this process, the volumetric CT data are transformed from an unstructured array of voxel intensities into a structured, hierarchical graph, where nodes represent coherent anatomical regions and edges encode spatial or functional relationships among them. This representation allows the learning model to reason explicitly about the topology of anatomical structures, enabling the discovery of relational patterns that are difficult to capture in traditional convolutional or transformer-based feature spaces [11,31].

3.3 GRAPH CONSTRUCTION: DEFINING NODAL AND EDGE ATTRIBUTES

Following the supervoxelization process, the next stage involves constructing a formal graph representation of the CT volume that captures the geometric and radiological relationships among the identified anatomical regions. The resulting structure is modeled as an undirected, weighted graph that serves as the primary input to the subsequent graph neural network (GNN). This graph encodes both local spatial adjacency and contextual intensity similarity, thereby allowing the model to reason over anatomical topology and tissue relationships [1, 27].

The set of vertices in this graph corresponds directly to the supervoxels obtained from the

Felzenszwalb segmentation. Each vertex represents a single, contiguous region of the CT volume, effectively an atomic structural unit within the larger anatomical context. The feature vector associated with each node encapsulates descriptive information about that supervoxel, such as its mean intensity, shape descriptors, and spatial coordinates. These nodal attributes provide the GNN with a compact yet semantically rich representation of each anatomical entity. The precise composition and normalization of these feature vectors are discussed in detail in the following chapter.

The edges of the graph define the connectivity between supervoxels and are established according to a hybrid criterion that balances geometric adjacency and appearance similarity. Two nodes are connected if their corresponding supervoxels are spatially proximate, specifically, if the Euclidean distance between their centroids lies below a predefined empirical threshold. This criterion ensures that the constructed graph faithfully captures the local topology of the anatomy, linking regions that are genuinely adjacent within the three-dimensional spatial domain of the scan.

To further enhance the representational power of the graph, each edge is assigned a weight that quantifies the degree of relationship between the two connected supervoxels. This weight is computed as a function of both their geometric proximity and their intensity-based similarity. Conceptually, the weight reflects how likely two supervoxels are to belong to the same anatomical structure. A common formulation expresses this relationship as a weighted sum of two terms: the spatial distance between the supervoxel centroids and the dissimilarity in their mean Hounsfield Unit (HU) intensities. Two scalar coefficients, denoted as balance parameters, control the relative influence of spatial and intensity factors. When the coefficient for spatial distance is larger, the model emphasizes geometric adjacency; conversely, when the coefficient for intensity similarity dominates, the model becomes more sensitive to radiodensity continuity across regions.

This weighted edge formulation provides the GNN with a nuanced structural context. It allows the network to distinguish, for example, between two neighboring regions that are spatially adjacent but belong to different organs, and those that are contiguous both spatially and radiologically, likely parts of the same tissue. By encoding both geometry and appearance into the graph's edges, the representation becomes capable of expressing complex, non-local dependencies that are fundamental to accurate organ and tumor delineation in medical imaging [1, 27].

In summary, this graph construction stage transforms the segmented CT data into a structured, relational format in which each node represents an anatomically coherent region, and each edge encodes a quantified relationship between such regions. This representation provides a principled foundation for the subsequent GNN-based segmentation model, which will operate on these node and edge attributes to infer high-level structural patterns and predict fine-grained segmentation labels with contextual awareness.

Chapter 4

ARCHITECTURE AND METHODOLOGY

4.1 THE HYBRID GNN–CNN PARADIGM

The proposed Medically Annotate Anything (MAA) framework introduces a novel hybrid architecture that strategically integrates the representational strengths of Graph Neural Networks (GNNs) and Convolutional Neural Networks (CNNs) to overcome the inherent limitations of monolithic deep learning models such as conventional CNNs or Transformers. Rather than relying on a single architectural paradigm, MAA adopts a multi-stage, multi-representational approach, enabling it to process medical images in both non-Euclidean graph space and Euclidean voxel space [31]. This design facilitates a more holistic understanding of pancreatic anatomy, allowing the model to reason about global contextual structures while simultaneously preserving fine-grained spatial precision essential for accurate tumor boundary delineation [10,23].

The architectural rationale of MAA rests on a deliberate division of labor between its two core components. The Graph Neural Network (GNN) backbone is responsible for coarse segmentation and high-level structural reasoning, operating in a graph-structured latent space that captures the relational and topological dependencies among image regions or superpixels. By representing the pancreas and its surrounding anatomical context as a graph, the GNN can model long-range spatial relationships and anatomical variability that are difficult to encode through local convolutions alone [1, 7]. This is particularly advantageous for pancreatic tumor segmentation, where understanding the global spatial arrangement, for instance, the relationship between the tumor, pancreatic duct, and nearby vasculature, is critical for accurate localization. The GNN thereby acts as a global context encoder, learning to infer broad structural cues and propagate information across spatially distant but semantically related regions [22,28].

Once the GNN has produced an initial coarse segmentation map and a globally coherent feature representation, this output is passed to the Convolutional Neural Network (CNN) refinement head, which operates in the Euclidean voxel domain. The CNN specializes in fine-grained boundary recovery, texture refinement, and high-frequency detail reconstruction [30]. Through its hierarchical convolutional filters and skip connections, the CNN refines the coarse predictions generated by the GNN, aligning them precisely with the underlying voxel-level intensity gradients. This ensures that the final segmentation is not only globally consistent but also locally precise, capturing intricate tumor morphologies, irregular edges, and subtle contrast transitions that are often blurred in graph-based outputs [12,26]

In principle, this hybrid GNN-CNN approach draws on the complementary advantages and disadvantages of the components that make it up. CNNs are inherently suitable for local feature extraction and dense spatial prediction, but modeling long-range dependencies with their fixed receptive fields is difficult [45]. GNNs, on the other hand, use data represented as nodes and edges within a graph, which makes them suited for representing global structure and relational dependencies, but GNN predictions often lack spatial precision and

lose pixel-level detail during message propagation [1, 7]. Combining these two paradigms, therefore, MAA takes on the challenge of establishing a balance between contextual understanding and spatial accuracy, merging implicated reasoning on a global scale with local details refinement [25, 31].

In addition, this modular design promotes a high degree of flexibility and interpretability. The separation of GNN and CNN components allows for independent optimization and interpretation of global and local reasoning processes, which is particularly useful when explainability and traceability of model developed decisions are paramount in the clinical setting. The graph-based representation also adds a reasonable structure for encoding multi-modal information, such as anatomical priors or a radiomic features as node attributes or edge weights, in the same representation [37,46].

A comparative analysis of CNNs, Transformers, and GNNs, highlighting their respective strengths, limitations, and suitability for pancreatic tumor segmentation, is presented in Table 3. This analysis reinforces the motivation behind the MAA framework's design, demonstrating how the synergistic integration of graph-based and convolution-based reasoning can yield a more robust, data-efficient, and anatomically informed segmentation pipeline capable of addressing the complex challenges inherent in pancreatic cancer imaging [1, 7, 45].

Architecture	**Primary Strength**	**Primary Weakness**	**Data Requirement**	**Suitability for Pancreatic Segmentation**
CNNs (e.g., U-Net)	Strong local feature extraction; excellent for boundary details. Inductive biases lead to data efficiency. [7]	Limited receptive field; poor at modeling long-range, global dependencies. [7, 12]	Moderate. Effective on smaller datasets due to strong inductive biases.	Good for local boundary refinement but may fail to capture global anatomical context, leading to localization errors.
Transformers (e.g., TransUNet)	Capturing global context and long-range dependencies via self-attention. [13, 14]	Data-hungry; prone to overfitting on small datasets. Can lose fine-grained localization details. [7, 16]	Very High. Weaker inductive biases require large-scale pre-training for optimal performance.	Potentially powerful for contextual understanding but challenged by the scarcity of annotated medical data, risking poor generalization.
GNNs (e.g., Graph U-Net)	Explicitly models relationships and non-	Requires explicit graph construction; can be computationally	Low to Moderate. Inductive biases are well-suited	Well-suited for modeling the structural relationships

	Euclidean structures. More robust on irregular geometry. [7]	complex. [7]	for relational data and smaller datasets. [7]	between anatomical components (represented as nodes), but may lack voxel-level precision.

Table 3: Comparative Analysis of Deep Learning Architectures for Medical Segmentation

4.2 STAGE 1: FROM VOXEL SPACE TO GRAPH DOMAIN VIA SUPERVOXELIZATION

The first stage of the Medically Annotate Anything (MAA) pipeline represents a crucial data transformation phase, in which the raw 3D Computed Tomography (CT) volume, originally represented in a dense, Euclidean voxel space, is systematically converted into a sparse, non-Euclidean graph representation. This transformation is essential for enabling the application of Graph Neural Network (GNN) architectures, which are designed to operate on relational data structures composed of nodes and edges rather than grid-like arrays. By transitioning from a voxel-based to a graph-based representation, the pipeline is able to model spatial and contextual relationships between anatomical regions in a manner that more naturally reflects the complex topological organization of human anatomy, particularly within the pancreas and surrounding structures [1, 27].

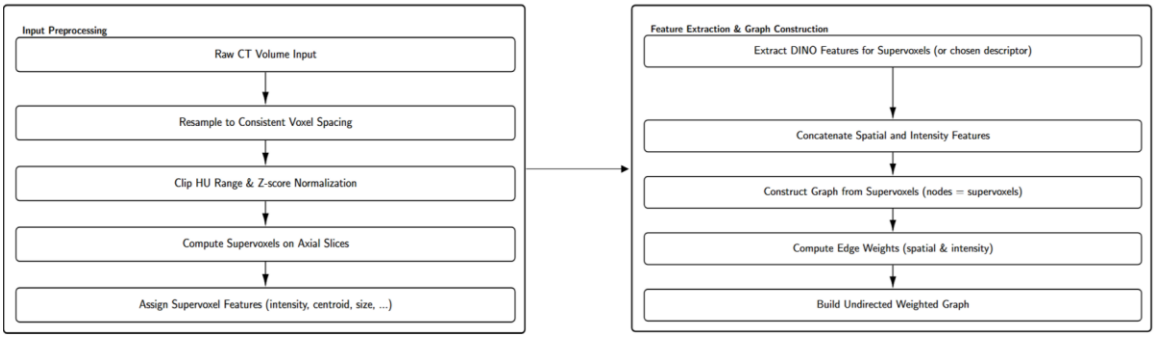

Figure 2: Workflow diagram of Input Preprocessing and Feature Extraction & Graph Construction

4.2.1) PREPROCESSING

The transformation process begins with a rigorous standardization of the input CT volumes to mitigate inter-scanner and inter-institutional variability. Variations in imaging acquisition protocols—such as slice thickness, field of view, and intensity calibration, can introduce significant heterogeneity in medical imaging data, which in turn degrades model generalization. To address this, all CT volumes are resampled to a uniform voxel spacing of 1.5*1.5*2 mm, ensuring spatial consistency across the dataset.

Next, intensity normalization is performed by clipping voxel intensities to a clinically relevant Hounsfield Unit (HU) range of [-125, 275], which corresponds approximately to soft tissue values encompassing the pancreas and surrounding abdominal organs. This step eliminates irrelevant high-density (bone) and low-density (air) regions, focusing the model's attention on diagnostically pertinent tissue. The clipped intensity values are

subsequently normalized via z-score normalization, which standardizes the data distribution and stabilizes model training. Collectively, these preprocessing steps serve to reduce domain shift and improve the robustness of downstream graph-based feature extraction and learning [17].

4.2.2) SUPERVOXEL GENERATION

Following preprocessing, the CT volume undergoes supervoxelization, a process that reduces the image from millions of individual voxels to a smaller number of perceptually coherent, spatially contiguous regions known as supervoxels. This is accomplished using the Felzenszwalb and Huttenlocher algorithm [11], a well-established graph-based image segmentation technique that efficiently groups neighboring pixels (or voxels) with similar intensity and texture characteristics.

In practice, the algorithm is applied on a slice-wise basis to the 2D axial slices of the CT volume, producing superpixels that are subsequently aggregated into 3D supervoxels. These supervoxels act as atomic visual units, each representing a compact and semantically meaningful region within the volume—such as a portion of pancreatic parenchyma, vasculature, or tumor tissue. By abstracting the image in this way, the computational burden of subsequent graph operations is dramatically reduced: instead of processing millions of voxels, the model now operates on a few thousand supervoxels. This reduction not only improves computational feasibility but also enhances semantic coherence, as the nodes correspond to regions with uniform appearance and anatomical context rather than individual noisy voxels.

4.2.3) NODE FEATURE EXTRACTION

Once supervoxels are defined, the next step involves assigning a rich feature representation to each node in the graph. Each node (supervoxel) must encapsulate both appearance-based and spatial information to enable effective graph-based reasoning. To achieve this, the MAA framework employs DINO (self-DIstillation with NO labels) [33], a self-supervised Vision Transformer (ViT) pretrained on large-scale image datasets, to extract high-level semantic embeddings from each supervoxel [6,14]

We select DINO for its potential to learn context-sensitive and semantically discriminative features without explicit supervision. Importantly, DINO is trained on natural image corpora, and its learned representations are transferrable to medical imaging tasks across domains (specifically soft-tissue imaging). To create a feature vector containing higher-order abstraction of detailed cues or features related to groupings of multiple voxels, we will pass each supervoxel through the pretrained DINO encoder. In addition, DINO-derived embeddings will be concatenated with other features to strengthen representational completeness. These features will include the normalized coordinates of the centroid of the supervoxel (to maintain spatial location in the 3D volume), as well as the mean intensity value for each supervoxel (to infer average tissue density). After this preparatory work, a multi-dimensional feature vector is formed by the inclusion of these embedding, ensuring each node of graph developed has sufficient semantic richness and geometric content, forming a strong basis for message passing and relational reasoning during the GNN stage of this analysis.

4.2.4) GRAPH CONSTRUCTION

The final step in this stage involves constructing the graph structure itself. Each supervoxel becomes a node v_i in V, and edges e_{ij} in E are established to represent meaningful relationships between nodes. The connectivity between nodes is determined by a weighted combination of spatial adjacency and feature similarity. Specifically, edges are formed between supervoxels that are spatially proximate (within a predefined Euclidean distance threshold) and exhibit similar intensity or feature embeddings, as measured by cosine similarity or Euclidean distance in the feature space [27].

The resulting undirected, weighted graph encodes both spatial topology and contextual similarity, providing an abstract yet anatomically grounded representation of the pancreas and its surrounding structures. Edge weights quantify the degree of relationship between regions, allowing the subsequent GNN to selectively propagate information along the most relevant connections. This graph-based abstraction is thus not merely a computational convenience—it constitutes a conceptual re-encoding of medical imagery into a form that better aligns with the relational and hierarchical nature of human anatomy. By representing the image as a graph of semantically meaningful regions rather than isolated pixels, the MAA framework sets the foundation for global reasoning about structural context and spatial relationships, which is indispensable for accurate and clinically reliable pancreatic tumor segmentation [1, 27].

4.3 THE GNN U-NET BACKBONE FOR COARSE STRUCTURAL SEGMENTATION AND CNN HEADS

At the heart of the Medically Annotate Anything (MAA) framework is a GNN-based U-Net backbone, which is the critical component used to perform coarse structural segmentation in the graph domain, after which structure extraction from the graph domain uses a convolutional-based reconstruction. By leveraging the supervoxel graph produced in Stage 1, the GNN U-Net permits the model to reason about spatial relationships in a non-Euclidean space, a vital capability when segmenting anatomically complicated organs, such as the pancreas, where relational dependencies and global context are important when delineating tumor boundaries [12].

Overall, the model resembles the traditional U-Net architecture [30] that has been extremely successful in convolutional segmentation tasks. The architecture retains the familiar U-net shape with an encoder-decoder topology and skip connections, but is reconceptualized for graph-structured data rather than grid-based feature maps. In this graph-based structure, the encoder extracts the input graph into coarser abstractions with increasingly meaningful semantic structures, while the decoder produces the abstractions into a segmentation map matching the topology of the original graph. The skip connections preserve essential low-level structural information by directly linking corresponding encoder and decoder stages, ensuring that fine-grained details are not lost during downsampling [12].

This design confers two critical advantages. First, it allows the network to capture hierarchical context, enabling deeper graph layers to integrate information across increasingly distant anatomical regions. Second, it ensures precise localization by fusing

these global abstractions with local structural cues during the decoding process. Together, these properties make the GNN U-Net particularly well-suited for coarse anatomical segmentation, where maintaining both contextual integrity and boundary awareness is essential before refinement in voxel space.

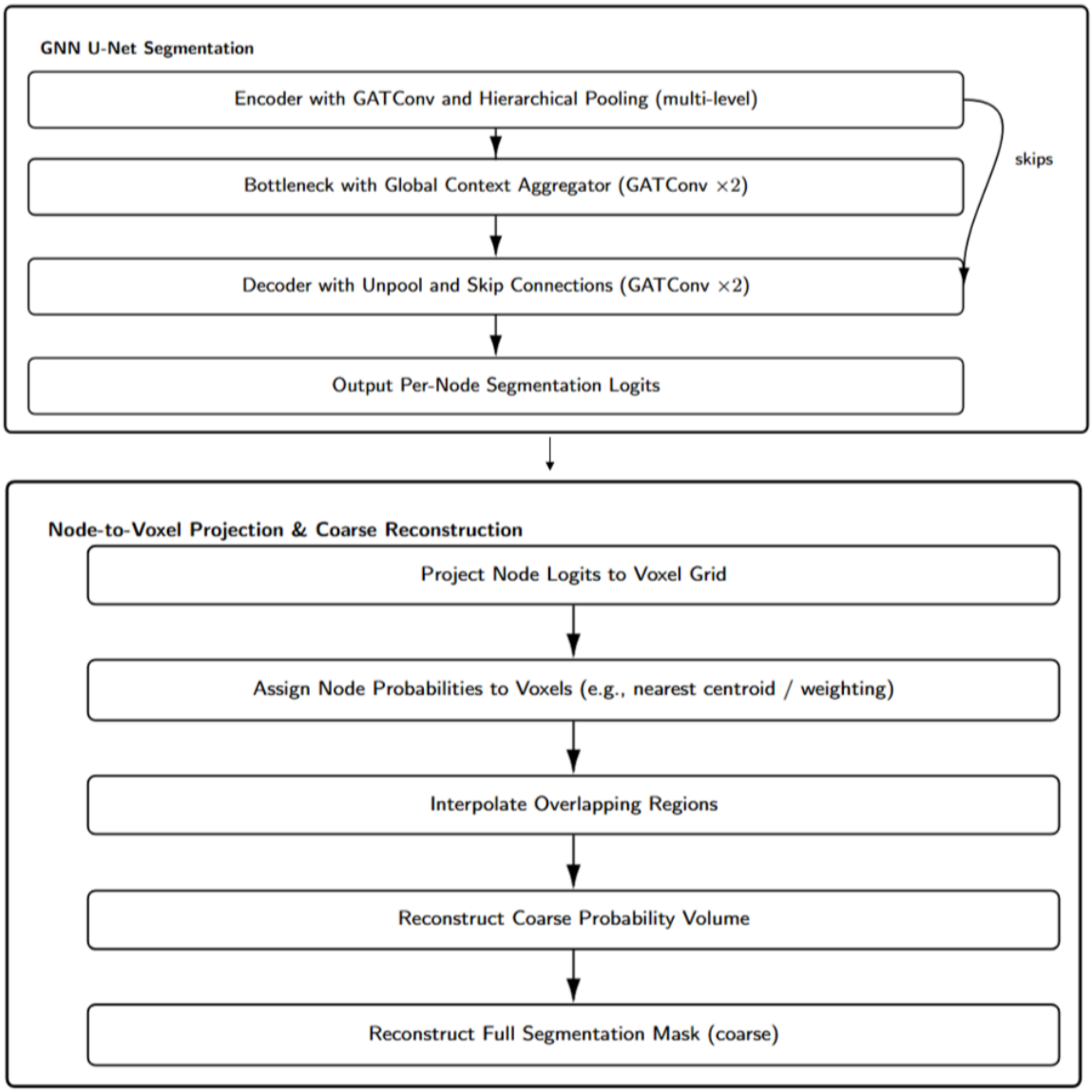

Figure 3: Workflow diagram of GNN UNet and Node to Voxel Projection and Reconstruction

4.3.1) ENCODER PATH: MULTI-SCALE GRAPH ABSTRACTION WITH GRAPH ATTENTION LAYERS

The encoder path of the GNN U-Net is designed to perform multi-scale abstraction of the input supervoxel graph by learning progressively higher-order feature representations. Each encoder stage consists of multiple Graph Attention Convolution (GATConv) layers [38], which collectively enable the model to adaptively aggregate information from each node's neighborhood. The use of GAT layers is a deliberate architectural choice due to their ability to learn data-driven, context-sensitive relationships between neighboring

nodes, something that conventional Graph Convolutional Networks (GCNs) cannot achieve effectively. In traditional GCNs, information propagation between connected nodes is governed by fixed weights derived from the graph's adjacency structure. While effective for capturing general connectivity patterns, this approach fails to account for heterogeneous importance among neighboring nodes. In contrast, Graph Attention Networks (GATs) introduce a self-attention mechanism that allows the model to learn how much influence each neighbor should have when updating a node's feature representation [38]. In practical terms, for every node in the graph, the GAT layer computes an attention coefficient for each of its neighbors based on their current feature representations. These coefficients quantify the relative importance of neighboring nodes and determine the extent to which their features contribute to the node's updated embedding.

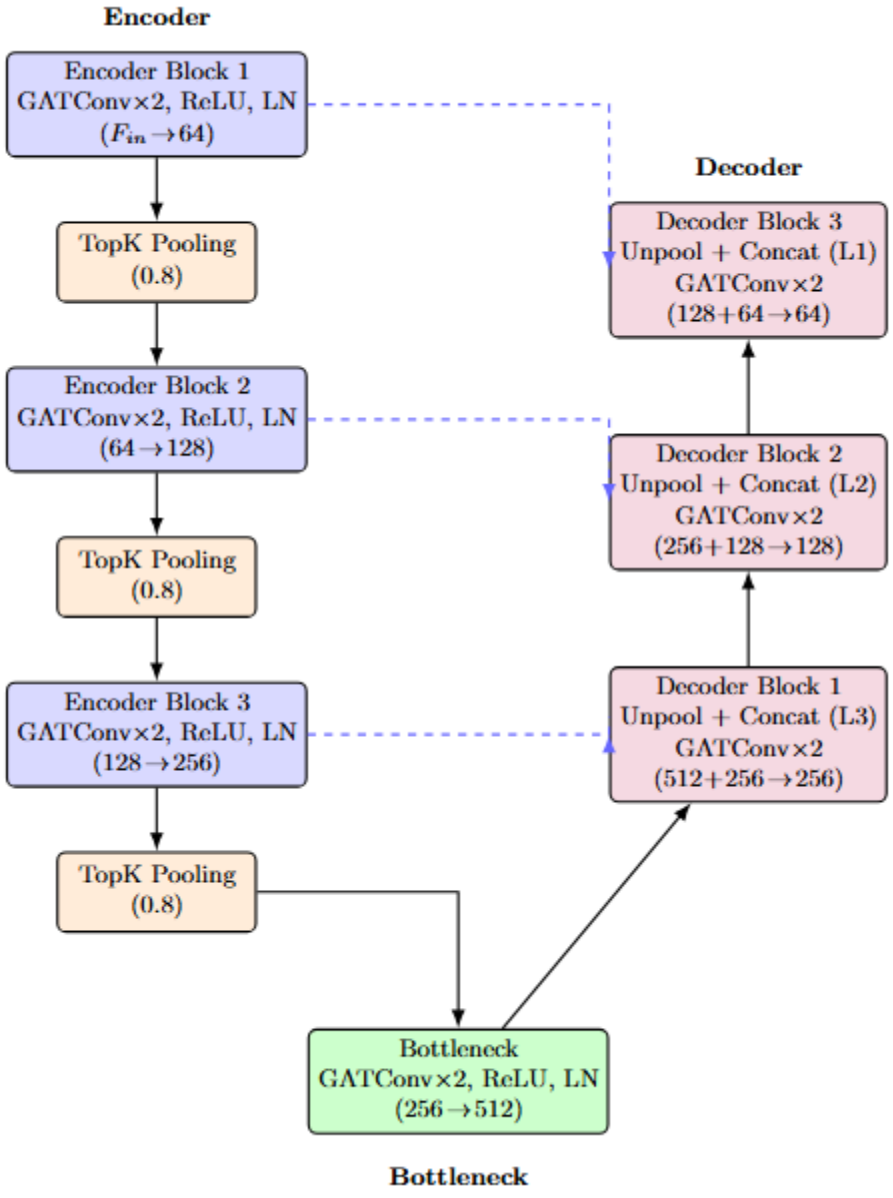

Figure 4: Architecture of the GNN U-Net Backbone

The new feature representation for a given node is thus computed as a weighted combination of its own features and those of its neighbors, where the weights correspond to the learned attention scores. These scores are normalized using a softmax function across all neighbors to ensure that the attention values form a valid probability distribution. Through this adaptive weighting process, the model learns to focus selectively on the most relevant neighbors, those that are anatomically or semantically significant for the segmentation task. For example, a supervoxel corresponding to pancreatic tissue might place higher attention on adjacent supervoxels representing tumor margins or vascular structures, while down-weighting less relevant regions such as surrounding fat tissue. To further enhance representational power and stabilize the learning process, the GNN employs multi-head attention, wherein multiple independent attention mechanisms are computed in parallel. Each attention head captures a distinct relational pattern or semantic dependency, and their outputs are concatenated or averaged to form the final node representation. This multi-headed design allows the encoder to model complex, heterogeneous relationships across the graph, thereby improving robustness and expressiveness [38].

As the graph is passed through successive encoder blocks, the model performs hierarchical pooling operations to gradually reduce the number of nodes and coarsen the graph [12, 41]. This process is conceptually analogous to spatial downsampling in CNN-based U-Nets, where resolution is traded for broader contextual awareness. In the graph domain, pooling can be achieved through algorithms such as top-k pooling or differentiable graph coarsening, which retain the most informative nodes based on learned importance scores while preserving overall structural integrity. The outcome of this hierarchical encoding is a compact, semantically enriched representation of the pancreas and its surrounding anatomy, which serves as the input for the subsequent decoding and refinement stages. In summary, the encoder path of the GNN U-Net transforms the input supervoxel graph into a progressively abstract and context-aware latent representation. By combining the adaptive attention capabilities of GAT layers [38] with the multi-scale abstraction afforded by hierarchical pooling [12], it enables the network to capture both local dependencies and long-range anatomical relationships—a critical requirement for effective coarse segmentation in complex organs such as the pancreas.

4.3.2) HIERARCHICAL DOWNSAMPLING VIA TOPK GRAPH POOLING

In order to promote multi-scale feature learning and hierarchical abstraction in the graph domain, the encoder path of the GNN U-Net architecture incorporates graph pooling layers that systematically reduce the graph's complexity [41]. This reduction process is analogous to a max-pooling operation in traditional CNN architectures, where the spatial resolution is reduced to focus upon the more semantically meaningful features. TopK Pooling is explored as a graph downsampling method within the proposed architecture because it enables an adaptive approach to retain the most informative subset of nodes and discard redundant or less useful nodes.

As a node selection method, TopK Pooling encompasses a learnable node selection mechanism, and therefore differs from a traditional pooling mechanism. TopK Pooling begins by learning a projection vector which describes a trainable filter used to evaluate the importance of a node in the graph. The feature of each node is projected onto this learned vector and provides an importance score [12] as a measure of how significant a node is to the task downstream, based on the nodes feature content and its structural context in the graph. A node with higher importance scores indicates it possesses more importance, and it is reasonable to expect it will represent an area of relevance (whether anatomical or pathological) for medical imaging tasks, such as tumor segmentation [1, 31].
Once importance scores have been computed, the nodes are ranked in descending order of relevance, and only a fraction of the most significant nodes, determined by a pooling ratio parameter, are retained. For example, if the pooling ratio is set to 0.5, only the top 50% of nodes with the highest importance scores are preserved. The remaining nodes are discarded, effectively yielding a coarser and more abstract graph that captures higher-level relationships between anatomical structures. Importantly, this process is not merely a form of dimensionality reduction; it also acts as an attention-driven filtering mechanism, focusing computational resources on the most diagnostically relevant supervoxels. The features of the retained nodes are scaled (or "gated") by their respective importance scores to modulate their influence during further message passing. A new adjacency matrix is then induced for the reduced set of nodes, ensuring that spatial and relational consistency is maintained in the coarsened graph [12].

By repeating this process across multiple levels of the encoder, the model constructs a hierarchical representation of the graph, transitioning from fine-grained, local structural details at the shallow layers to global, semantically abstract relationships at deeper layers [41]. This hierarchical feature encoding is crucial in medical segmentation tasks, where the distinction between tumor boundaries and surrounding tissue often depends on both localized intensity patterns and broader contextual cues. In essence, TopK Pooling provides the GNN backbone with the capacity to learn structurally aware, scale-invariant representations that enhance its ability to perform coarse segmentation and structural reasoning in the complex, non-Euclidean topology of medical data.

4.3.3) DECODER PATH: GRAPH UNPOOLING AND FEATURE PROJECTION

The decoder path of the GNN U-Net serves as the counterpart to the encoder and is responsible for reconstructing a detailed, high-resolution segmentation map from the abstracted, low-resolution graph representations produced by the encoder. Its primary goal is to reverse the hierarchical compression that occurred during encoding, progressively recovering both spatial and structural details while preserving the semantic richness of the learned features [12].

The decoder follows a mirrored U-shaped structure, composed of a sequence of blocks that each perform two main operations: graph unpooling and graph attention-based convolution (GATConv) [38]. The unpooling operation reverses the effects of the earlier pooling steps, restoring the graph to the resolution it had at a corresponding encoder stage. This is typically achieved by routing the node features from the coarsened graph back to their

original node indices or positions recorded during pooling. In essence, unpooling acts as a mapping function that reassigns high-level feature vectors to the appropriate nodes in the higher-resolution graph, thus recovering fine structural granularity [12].

After unpooling, each decoder block applies one or more Graph Attention Convolution (GAT) layers [38]. These layers refine the upsampled feature representations by aggregating contextual information from neighboring nodes, guided by learned attention coefficients. In this phase, the GAT layers play a critical role in reintroducing local structural coherence that may have been lost during pooling. They allow the network to propagate semantic information effectively across the reconstructed graph, ensuring that the final node features are contextually consistent and spatially aligned with anatomical boundaries.

One important architectural advancement that improves the representational capacity of the decoder is the inclusion of skip connections. At each hierarchical level, the decoder's feature maps are concatenated with those of the encoder feature maps at the same resolution. This direct fusion of features integrates high-level semantic understanding (from deeper layers) with fine-grained spatial detail (from shallower layers) [12]. In practice, the use of skip connections prevents losing boundary sharpness and texture information typically found in models that rely solely on deep and abstracted features. This design emulates the rationale of original CNN-based U-Net [30] being highly successful in biomedical segmentation applications and enables for a balance to be struck within the GNN U-Net framework for globally understanding context with local details retained.

The last component of the decoder is the GATConv output layer, which produces prediction logits that are per-node. Each node's outputs infer the probability or confidence that the associated supervoxel is a specific class, such as tumor or background tissue. The outputs of these coarse predictions serve as the baselines for the segmentation map in the graph domain. In the overall Medically Annotate Anything (MAA) framework, this graph-based coarse map is subsequently refined in the voxel space by the CNN head to recover sharp tumor boundaries and local texture continuity.
In summary, the decoder path performs a structured and information-rich reconstruction process. Through hierarchical graph unpooling [12], attention-guided feature propagation [38], and cross-scale fusion via skip connections [12, 30], it enables the GNN U-Net backbone to produce a segmentation that captures both the global anatomical structure and the fine morphological details of pancreatic tumors, achieving a level of interpretability and precision that purely Euclidean convolutional models struggle to match.

Level	Component	Operation Details	Input Channels	Output Channels	GAT Heads	Pooling Ratio
1	Encoder Block 1	GATConv × 2, ReLU, LayerNorm	Fin	64	8	-
	Pooling 1	TopK Pooling	64	64	-	0.8
2	Encoder Block 2	GATConv × 2, ReLU, LayerNorm	64	128	8	-
	Pooling 2	TopK Pooling	128	128	-	0.8

3	Encoder Block 3	GATConv × 2, ReLU, LayerNorm	128	256	8	-
	Pooling 3	TopK Pooling	256	256	-	0.8
4	Bottleneck	GATConv × 2, ReLU, LayerNorm	256	512	8	-
5	Decoder Block 1	Unpool, Concat (from Level 3), GATConv × 2	512 + 256	256	8	-
6	Decoder Block 2	Unpool, Concat (from Level 2), GATConv × 2	256 + 128	128	8	-
7	Decoder Block 3	Unpool, Concat (from Level 1), GATConv × 2	128 + 64	64	8	-

Table 4: Architectural Specifications of the GNN U-Net Backbone - This table provides a hypothetical yet representative layer-by-layer configuration for the GNN U-Net, detailing the operations, feature dimensions, number of attention heads, and pooling ratios at each level. Fin refers to the dimension of the initial node features, and $N_{classes}$ represents the number of segmentation classes. This specification is crucial for ensuring reproducibility and understanding the model's overall capacity.

4.3.4) REPROJECTION FROM THE GRAPH DOMAIN TO THE VOXEL SPACE

After the GNN U-Net backbone processes the graph and generates per-node segmentation logits, these predictions need to be mapped back from the abstract graph domain to the original three-dimensional voxel space of the CT scan. This reprojection step acts as the bridge between the coarse, relational segmentation produced by the GNN and the fine-grained spatial refinement that follows [31].

The procedure is conceptually simple yet essential. Each node in the graph corresponds to a particular supervoxel in the 3D CT volume. The segmentation probability vector, or logits, predicted for that node is assigned uniformly to all voxels belonging to its corresponding supervoxel. In effect, this operation "paints" the predicted label probabilities back onto the voxel grid according to the supervoxel boundaries. The outcome of this reprojection is a coarse probability volume that matches the dimensions of the original CT scan. However, this volume appears somewhat "blocky" or discretized, since all voxels within a supervoxel share identical probability values. Despite the coarse spatial granularity at the edges, this representation preserves the global context and anatomical consistency captured by the GNN. Consequently, it serves as a robust and contextually informed foundation for the subsequent CNN-based refinement stage, which will recover fine structural details and sharpen tumor boundaries.

4.3.5) THE 3D CONVOLUTIONAL REFINEMENT HEAD FOR BOUNDARY DELINEATION

The reprojection of supervoxel-level predictions from the GNN naturally produces coarse and blocky segmentation boundaries that do not perfectly align with the true anatomical contours of the pancreas or the tumor. To overcome this limitation, the framework introduces a dedicated 3D Convolutional Neural Network (CNN) Refinement Head, specifically designed to enhance boundary precision and restore fine structural details lost

during the earlier abstraction and supervoxelization stages [31]. Rather than processing the entire CT volume, the refinement head operates on a cropped subregion, a tightly bounded area around the region of interest identified by the GNN's coarse segmentation. This targeted strategy significantly reduces computational overhead while allowing the CNN to focus exclusively on the most diagnostically relevant area, thereby improving both efficiency and accuracy. The input to the refinement head is a multi-channel 3D tensor, composed of several layers of information stacked together to provide rich contextual and structural cues:

- The original preprocessed CT volume within the cropped region (1 channel).
- The coarse probability volume generated by the GNN, representing the per-voxel segmentation probabilities for each class (C channels, where C is the number of segmentation classes, in our work is 28).
- Optionally, reprojected node feature maps, derived from the graph embeddings and mapped back into voxel space (F channels, where F is the feature dimension).

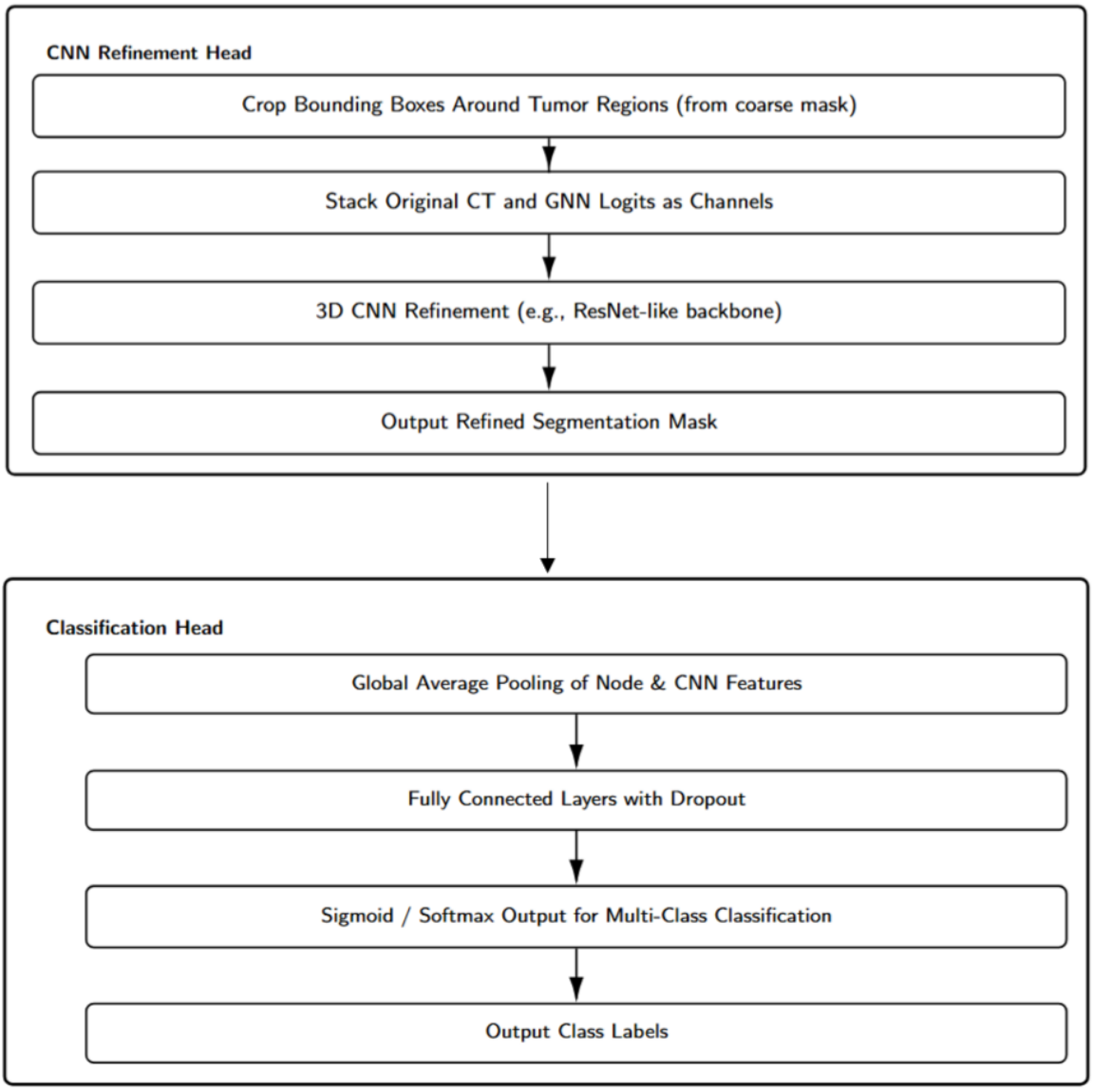

Figure 5: Workflow diagram of 3D CNN Refinement and Classification

The refinement network itself is built as a lightweight 3D CNN, which is a modified ResNet architecture optimized for volumetric data. This design leverages the CNN's strong inductive bias toward learning local spatial patterns, enabling it to detect subtle intensity gradients and texture variations that delineate precise tumor boundaries. Through a series of 3D convolutional layers, nonlinear activations, and normalization operations, the network progressively sharpens boundary transitions and corrects the coarse, supervoxel-induced artifacts from the GNN output.

Conceptually, this GNN–CNN hybrid structure embodies a complementary partnership between two distinct computational paradigms [10, 25, 31]. The GNN, operating in the graph domain, performs global localization and structural reasoning, capturing the broader anatomical context and spatial relationships between organs and tissues. The CNN, on the other hand, specializes in local refinement, exploiting its convolutional locality to recover fine-grained details and exact edge contours.
This design establishes a division of labor:

- The GNN identifies where the tumor is likely located, providing a coarse but contextually rich spatial prior.
- The CNN then uses that prior as guidance to determine exactly where the tumor boundaries lie, refining the segmentation with high spatial precision.

In effect, the GN provides the global "what and where," while the CNN provides the local "how exactly." This symbiotic interaction between global reasoning and local detail recovery allows the system to achieve both semantic consistency and boundary-level accuracy, producing segmentation results that are both anatomically coherent and clinically actionable.

4.3.6) THE DIAGNOSTIC CLASSIFICATION HEAD

Not only does the framework accomplish segmentation, but it can also be used for diagnostic purposes at the patient level, a key step in establishing a more complete, end-to-end clinical decision support system [1]. This diagnostic part of the framework is called the Classification Head and provides predictions of higher-level clinical properties, such as tumor type, malignancy grade, or disease stage. By bringing together information from both the graph-based structural reasoning, and the voxel-based spatial refinement, the classification head allows the model to generalize from solely spatial segmentation to its clinical meaning.
The classification head operates on a compact, global feature representation that summarizes the entire predicted tumor region. This feature vector serves as a condensed descriptor that integrates both macrostructural and microstructural information captured at different stages of the pipeline. It is constructed through a multi-step feature aggregation process:

- Graph-based feature aggregation: From the GNN backbone, all node embeddings corresponding to supervoxels classified as part of the tumor are extracted. These embeddings encapsulate the high-level relational and structural characteristics of the tumor region within the anatomical context of the pancreas and surrounding

organs. To create a single representation, these node features are aggregated using a global pooling strategy, commonly global average pooling or global attention pooling, which effectively summarizes the distributed node-level information into one vector that represents the tumor as a coherent graph entity.

- Voxel-based feature aggregation: From the CNN refinement head, features from the final convolutional layer are extracted, focusing only on the voxels within the refined tumor mask. These features encode fine-grained textural and intensity-based information, such as edge sharpness, tissue heterogeneity, and internal tumor morphology, attributes that are often critical for clinical diagnosis. Similar to the graph domain, these voxel-level features are reduced to a fixed-length representation through spatial pooling.
- Feature fusion: The pooled feature vectors from the GNN and CNN are then concatenated to form a comprehensive, multi-modal descriptor. This fusion effectively combines the global contextual awareness of the GNN with the local detail sensitivity of the CNN. The resulting hybrid representation captures both the structural organization of the tumor within its anatomical neighborhood and the fine visual cues that characterize its internal texture and boundaries.

Once the unified feature vector is constructed, it is passed through a Multi-Layer Perceptron (MLP) that functions as the actual classifier. The MLP is composed of several fully connected layers interleaved with non-linear activation functions (such as ReLU) to enable complex decision boundaries. Dropout layers are included for regularization, reducing overfitting and improving the model's ability to generalize to unseen patient data. The final output layer employs either a sigmoid activation function (for binary classification tasks) or a softmax activation function (for multi-class or multi-label problems), producing probabilistic predictions that correspond to the diagnostic categories of interest—such as tumor subtype or clinical stage.

From a design perspective, the classification head serves as a natural culmination of the hybrid GNN–CNN architecture. While the segmentation pipeline ensures spatial accuracy and anatomical fidelity, the classification module distills that spatial understanding into a clinically interpretable diagnostic output [1]. This integration bridges the gap between computational perception and medical decision-making, enabling the system not only to delineate where the tumor is but also to infer what it is and how advanced it might be.

In essence, the diagnostic classification head transforms the framework from a purely image-analysis tool into a multi-functional diagnostic assistant—one capable of synthesizing complex spatial, structural, and textural cues into actionable clinical insights [1].

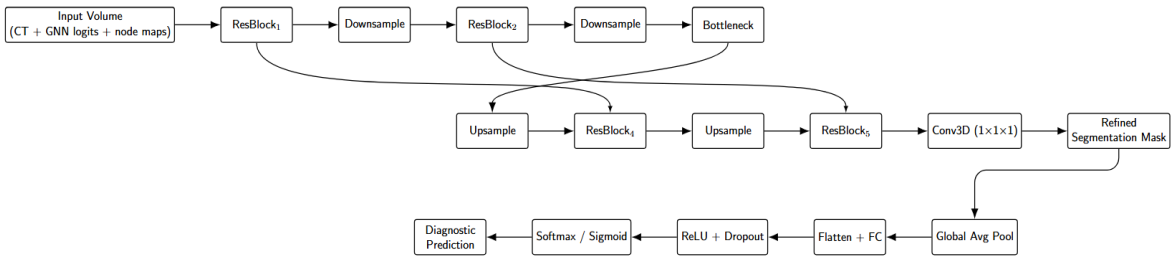

Figure 6: Architecture of the 3D CNN Refinement and Classification heads

4.4 POSTPROCESSING AND POST-TRAINING

Following the primary segmentation inference, several post-processing and post-training operations were applied to refine, stabilize, and optimize the final predictions. These stages ensure that the raw output from the neural network adheres to anatomical plausibility, removes spurious predictions, and enhances overall reliability during evaluation and deployment.

4.4.1) POST-PROCESSING PIPELINE

The post-processing module was designed to eliminate noise, enforce spatial coherence, and produce anatomically consistent segmentation masks. Given that deep learning models, especially graph-based and hybrid architectures, can occasionally yield fragmented or noisy segmentations, this step serves as a structural regularizer on the model's output.

The pipeline consisted of the following key operations:

- Keep Largest Connected Component: For each predicted class, only the largest connected component was retained. This ensures that the final mask corresponds to the primary anatomical structure rather than small, isolated false positives. This step is particularly important in organs such as the pancreas or gallbladder, where background textures or nearby tissues can produce small spurious activations.
- Remove Tiny Objects: Any predicted regions below a predefined voxel threshold were removed. This acts as a noise suppression filter, discarding predictions that are too small to represent meaningful anatomical entities. The threshold was empirically determined based on validation set statistics to balance sensitivity and specificity.
- Smooth Boundaries (Morphological Closing): A 3D morphological closing operation (dilation followed by erosion) was applied to each organ mask to remove small holes and discontinuities. This process smooths rough boundaries and enhances the topological continuity of organs, improving visual realism and metric performance, particularly for boundary-sensitive measures such as Hausdorff Distance.
- Calibrate Thresholds with Validation Data: The segmentation confidence thresholds were fine-tuned using the validation set. Instead of relying on a fixed 0.5 probability cutoff, optimal thresholds for each organ were determined to maximize the Dice score. This calibration step helps correct for class imbalance and accounts for varying model confidence across structures of different sizes.
- Output Cleaned Segmentation Mask and Labels: After all refinements, the pipeline produced a final cleaned segmentation mask along with corresponding organ labels. These outputs were stored in NIfTI format to ensure compatibility with standard medical imaging toolkits and subsequent quantitative evaluation workflows.

Through this structured post-processing pipeline, the raw model outputs were transformed into topologically coherent and anatomically faithful segmentations. Qualitative inspection confirmed smoother boundaries, reduced false positives, and improved inter-organ separation, all of which contributed to better overall quantitative performance.

4.4.2) POST-TRAINING MODULE

In addition to output refinement, a post-training module was implemented to further stabilize the model predictions and prepare the trained models for inference and deployment. This stage incorporated three complementary components: model ensembling, calibration, and export.

Model Ensembling: Multiple model checkpoints from the final training epochs were averaged using soft-voting ensemble strategies. The ensemble prediction was computed as the voxel-wise mean of class probability maps across selected models. This approach mitigates the stochastic variability of individual models, leading to smoother and more robust predictions. Empirically, the ensemble improved both Dice and HD95 scores, confirming that model diversity can enhance generalization.

Calibration of Model Confidence: To address overconfidence in neural network predictions, temperature scaling was applied to recalibrate the model's output probabilities. Proper calibration ensures that predicted probabilities better reflect true likelihoods, improving interpretability in clinical and decision-support contexts. The optimal temperature parameter was determined on the validation set by minimizing the negative log-likelihood.

Model Export and Deployment Readiness: The final calibrated and ensembled model was exported to an optimized format (e.g., TorchScript or ONNX) for deployment and reproducibility. All accompanying configuration files, normalization parameters, and class mappings were stored to facilitate consistent inference on new datasets. This stage ensures that the trained system is portable, reproducible, and compatible with downstream processing pipelines.

Together, the post-processing and post-training modules form a crucial bridge between raw model predictions and clinically meaningful segmentation outputs. Post-processing enforces structural and geometric validity, while post-training enhances predictive stability and interpretability. These refinements not only improve quantitative metrics but also ensure that the model outputs are visually consistent, anatomically plausible, and ready for integration into broader medical imaging workflows.

Chapter 5

LEARNING, INFERENCE, AND RESULTS

5.1 A MULTI-COMPONENT LOSS FUNCTION FOR HOLISTIC TRAINING

The training of the proposed hybrid GNN–CNN architecture presents unique challenges due to the model's modular structure and the diverse nature of its outputs. The GNN component operates on graph-structured representations to perform supervoxel-level reasoning and coarse segmentation, while the CNN refinement head focuses on voxel-level boundary delineation. Additionally, the diagnostic classification head outputs patient-level predictions. Consequently, a simple single-objective optimization function would be inadequate to balance these distinct yet interdependent learning objectives.

To address this, the model is trained end-to-end using a multi-component composite loss function, denoted as L_{total}, which integrates several complementary objectives. This allows the model to jointly optimize for segmentation accuracy, diagnostic classification performance, and structural consistency between the GNN and CNN representations. The overall objective is expressed as a weighted sum of the constituent losses:

$$L_{total} = \lambda_1 * L_{seg} + \lambda_2 * L_{cls} + \lambda_3 * L_{aux} \quad (5.1)$$

where $\lambda_1, \lambda_2, \lambda_3$ are non-negative weighting coefficients that control the relative contribution of each term. These weights can be tuned empirically based on the desired emphasis, whether toward segmentation fidelity, diagnostic accuracy, or representational coherence.

5.1.1) SEGMENTATION LOSS (L_{seg})

The segmentation objective drives the model to accurately delineate tumor regions within the input volume. It is composed of two synergistic loss components designed to address both regional accuracy and boundary precision:

(a) Voxel-wise Cross-Entropy Loss: This term constitutes the primary supervision signal for semantic segmentation. It measures the per-voxel divergence between the predicted probability distribution and the ground truth segmentation mask. Each voxel is treated as an independent classification problem, and misclassifications are penalized proportionally to their predicted confidence. This ensures that the CNN refinement head learns robust class separation across the volume.

(b) Boundary-Aware Loss: Tumor boundaries are often indistinct, and coarse supervoxel-based segmentation can introduce spatial imprecision. To explicitly encourage sharper delineation, a boundary loss is incorporated [21]. This may take several forms—such as a Dice loss computed on a narrow band of voxels surrounding the ground-truth boundary, or a loss based on the discrepancy between the signed distance transforms (SDTs) of the predicted and ground-truth masks. Both formulations incentivize the model to refine the interface between the tumor and surrounding tissue, thus improving anatomical plausibility.

The combined segmentation loss can therefore be expressed as:

$$L_{seg} = \alpha_1 * L_{CE} + \alpha_2 * L_{boundary} \quad (5.2)$$

where α_1 and α_2 control the internal balance between region-level and boundary-level supervision.

5.1.2) CLASSIFICATION LOSS (L_{cls})

For patient-level predictions such as tumor subtype or stage, the diagnostic classification head uses a Categorical Cross-Entropy Loss to supervise this task. The Categorical Cross-Entropy Loss calculates the difference between the predicted probability vector and the ground-truth class label and therefore, as a loss, directs the Multi-Layer Perceptron (MLP) classifier to consider information from both the graph embeddings and the CNN feature maps to arrive at a single diagnostic decision. Using a sigmoid activation for binary diagnosis tasks followed by binary cross-entropy or a softmax activation in multi-class settings followed by Categorical Cross-Entropy provides loss functions that ensure the model's final output integrates local appearance cues, and generalizes beyond the local cues to inform the global relational structure inferred by the GNN.

5.1.3) AUXILIARY AND REGULARIZATION LOSSES (L_{aux})

In addition to the primary objectives, several auxiliary losses are introduced to regularize intermediate representations and enforce cooperation between the GNN and CNN modules. These terms do not directly supervise the final outputs but provide structural constraints that enhance stability and generalization.

(a) Graph Regularization Loss:

Applied to the GNN output, this term promotes smoothness over the graph structure. It penalizes large differences in predictions between neighboring nodes with similar features, thereby ensuring spatial coherence. Formally, it can be expressed as a Laplacian regularization term [44] that minimizes $\sum_{(i,j)\in \mathrm{E}} \left|\left|y_i - y_j\right|\right|^2$ where E denotes the set of edges in the graph.

(b) Consistency Loss:

This is arguably the most critical auxiliary objective, serving as the coupling mechanism between the GNN and CNN domains. It enforces agreement between the coarse segmentation map predicted by the GNN in supervoxel space and the refined voxel-wise segmentation produced by the CNN. This can be realized through either Kullback–Leibler (KL) divergence or Mean Squared Error (MSE) computed between the corresponding probability maps. The effect is bidirectional: the CNN regularizes the GNN's coarse reasoning through local feedback, while the GNN provides a contextual prior to guide the CNN's refinement process [25, 39].

(c) Attention Regularization:

Within the GNN, attention coefficients computed by the GATConv layers determine how strongly each node attends to its neighbors [38]. To encourage interpretability and sparsity, an additional regularization term is introduced that penalizes diffuse attention distributions. This can be achieved by applying an entropy-based or L_1 penalty to the attention weights, forcing the model to focus on a smaller, more meaningful subset of neighboring nodes rather than uniformly averaging all connections.

This composite loss framework transforms the training process from a conventional supervised learning problem into a multi-objective optimization task governed by interdependent constraints. Each loss term acts as a specialized instructor, guiding a particular subsystem of the model. The segmentation losses train the CNN to refine spatial details [21, 47], the classification loss ensures global diagnostic reasoning, and the auxiliary terms maintain representational coherence and interpretability [44].

Most importantly, the Consistency Loss functions as the linchpin that unites the GNN's

structural intelligence with the CNN's local precision [25, 39]. Through this mechanism, the hybrid model evolves into a cohesive system in which both components learn to complement each other—bridging the gap between global relational reasoning and fine-grained boundary reconstruction.

5.2 POST-PROCESSING AND MASK FINALIZATION FOR CLINICAL APPLICABILITY

The raw voxel-wise probability maps generated through the CNN refinement head represent the model's direct pixel-wise predictions for tumor presence at every spatial location. However, these outputs in their raw form may consist of small spurious regions, jagged edges and artifacts that sometimes occur due to being noisy, imperfectly predicted or some uncertainty inherent in the data. While these artifacts are acceptable when evaluating an algorithm, they are not acceptable in a clinical context, where segmentation masks should be expected to be anatomically preserved, smooth and interpretable.

To address this, we implement a systematic post-processing pipeline to then take the raw CNN output to a final, deployable, clinically relevant segmentation mask [17]. This post-processing serves as an important bridge between algorithmic performance, and actual deployability to practice the final output can be expected from clinical morphology and diagnostic expectations.

5.2.1) THRESHOLDING AND BINARIZATION

The first step involves converting the continuous voxel-wise probability map into a binary segmentation mask. Each voxel is assigned to the tumor class if its predicted probability exceeds a fixed threshold—typically 0.5, although this value may be optimized empirically based on validation data. Formally, for each voxel v with predicted probability $p(v)$, the binary label $b(v)$ is defined as:

$$b(v) = 1 \text{ if } p(v) \geq 0.62, \text{ and } b(v) = 0 \text{ otherwise.} \quad (5.3)$$

This simple decision rule transforms the probabilistic output into a discrete representation, delineating the predicted tumor region. The thresholding step ensures interpretability while maintaining a direct correspondence with conventional binary ground-truth masks used in medical imaging datasets.

5.2.2) CONNECTED COMPONENT ANALYSIS

After thresholding, the binary mask typically contains several disconnected components, some of which will be small false-positive clusters or artifacts caused by noise. To remove the irrelevant components, we perform Connected Component Analysis (CCA).

CCA detects all contiguous 3D regions present in the binary mask (connected components). Each connected component is identified and labeled as an object. The connected components are then filtered or evaluated based on volume. In most clinical segmentation applications, such as pancreatic or hepatic tumors, it is reasonable to assume that the pathological region of interest is connected and represents a single mass. Therefore, the identified object is filtered by keeping the largest connected component and removing all smaller components.

This simple heuristic effectively removes false-positive isolated regions, which may be present because of similar intensities or ambiguous boundaries with surrounding tissues. The final product is a more contiguous and anatomically relevant segmentation. [17].

5.2.3) MINIMUM VOLUME FILTERING

To add further biological realism, we can apply a minimum volume constraint whereby we remove any predicted object that is below a threshold volume value predetermined to correspond with a tumor spine size that a clinician could reasonably estimate could be detected.

Thresholds can be determined using a subjective process that could be defined by the medical domain expert or from a recommendation applied to radiological evidence. For instance, a tumor size of less than a few cubic millimeters within an abdominal CT study would be indistinguishable from normal background or noise from the imaging study. Therefore, the removal of a predicted small tumor (and its classification as a tumor) allows for the model to abstain from predicting insignificant positive results in clinical interpretation.

5.2.4) MORPHOLOGICAL REFINEMENT

After extraneous features and/or predicted components are removed, the remaining segmentation mask could present uneven or jagged edges from voxel-level prediction noise and/or discretization. Again, this could be ameliorated with a morphological operation that better represents the structure of the mask. One operation for doing this is called closing or morphological closing, where an object is first dilated then eroded using a small structuring element, it is often spherical in shape. Dilation means that the boundaries of the object will expand slightly so that the dilation can fill gaps or holes in the mask. Erosion returns the object to a volume approximately equal to the original object while smoothing the boundaries of the object as well. This process serves multiple purposes:

- It fills small internal cavities that may result from imperfect CNN confidence.
- It smooths sharp corners and staircase-like artifacts introduced by voxelization.
- It yields a boundary more consistent with the expected smooth morphology of soft tissue tumors.

The parameters of these morphological operations, such as the radius of the structuring element, are chosen based on the voxel resolution of the input CT scans and the expected size of the target structure.

5.2.5) FINAL MASK GENERATION AND INTEGRATION

After applying these steps, the result is a finalized binary segmentation mask that is both topologically simple and anatomically realistic [17]. This mask can be directly overlaid on the CT scan for visualization, used for volumetric quantification, or integrated into downstream diagnostic workflows.

In the broader context of the framework, the refined segmentation mask complements the diagnostic output of the classification head. Together, they provide a comprehensive clinical inference: a spatially localized segmentation of the tumor and a corresponding high-level diagnostic prediction (e.g., tumor subtype or stage).

By enforcing morphological plausibility, spatial coherence, and interpretability, the post-processing pipeline transforms the CNN’s raw predictions into a clinically actionable segmentation output, suitable for use in computer-assisted diagnosis, surgical planning, or quantitative radiomic analysis.

5.3 PRIMARY EVALUATION METRICS

The quantitative evaluation of a medical image segmentation framework must capture both the volumetric accuracy of the predicted regions and the precision of their boundaries. In the context of tumor segmentation, it is not sufficient for a model to merely identify the general region of interest; it must also delineate its exact borders with anatomical fidelity. To this end, two complementary and widely recognized metrics are employed in this study: the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD). Together, these metrics provide a comprehensive assessment of segmentation quality across both volume and boundary domains [29].

5.3.1) DICE SIMILARITY COEFFICIENT (DSC)

The Dice Similarity Coefficient, often referred to as the Dice Score, is an overlap-based measure that quantifies the degree of spatial agreement between the predicted segmentation and the ground-truth annotation. Given two sets: X representing the predicted voxels and Y representing the ground-truth voxels, the Dice coefficient is defined as:

$$\mathrm{DSC} = \frac{2\times|X\cap Y|}{|X|+|Y|} \tag{5.4}$$

Here, $|X \cap Y|$ denotes the number of voxels correctly identified as belonging to the target region by both the prediction and the ground truth, while $|X|$ and $|Y|$ represent the total number of voxels in the predicted and reference regions, respectively.

The Dice coefficient ranges from 0 to 1, where 0 indicates no overlap and 1 indicates perfect overlap. In practical medical imaging scenarios, a Dice score above 0.8 is typically considered satisfactory, though acceptable thresholds can vary depending on the anatomy and imaging modality.

The principal advantage of DSC lies in its robustness to class imbalance, which is a common challenge in medical datasets where the tumor region occupies only a small fraction of the total image volume [29]. By normalizing the intersection by the combined sizes of both regions, the Dice metric provides a balanced measure that does not disproportionately penalize models for the small size of the foreground class. This makes it particularly suitable for tumor segmentation tasks, where even small absolute errors can represent significant relative discrepancies.

Moreover, the Dice coefficient correlates closely with clinical intuition — higher values indicate a greater degree of spatial congruence between automated and expert annotations, directly reflecting how accurately the algorithm captures the tumor's extent.

5.3.2) HAUSDORFF DISTANCE (HD)

While the Dice coefficient captures the overall overlap between two segmentations, it does not explicitly account for boundary precision. A segmentation may achieve a high DSC by correctly identifying the bulk of the tumor region while still exhibiting large local deviations along its boundaries [29, 34]. The Hausdorff Distance complements the Dice metric by measuring the degree of spatial misalignment between the boundaries of the predicted and reference segmentations.

Formally, given two sets of boundary points, A (predicted boundary) and B (ground-truth boundary), the Hausdorff Distance is defined as the maximum of two directed distances:

$$\mathrm{HD}(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B}|a-b|,\ \sup_{b\in B}\inf_{a\in A}|b-a|\right\} \tag{5.5}$$

In plain terms, this metric computes the maximum distance from a point on one surface to

the closest point on the other surface. Thus, it captures the worst-case deviation between the predicted and true boundaries. A lower HD value indicates better spatial alignment and, therefore, higher segmentation accuracy.

However, because the traditional Hausdorff Distance is highly sensitive to small outlier points (for example, single misclassified voxels far from the main region), a more stable variant known as the 95th Percentile Hausdorff Distance (HD95) is often used [29, 34]. The HD95 measures the distance below which 95% of the boundary points lie, effectively discounting extreme outliers and providing a more reliable and clinically interpretable metric.

This boundary-based evaluation is particularly important in medical applications where precise delineation of tumor margins is critical for surgical planning, radiation therapy, and longitudinal monitoring. Even small errors near the tumor boundary can lead to substantial differences in treatment outcomes [29].

5.3.3) Complementarity of the Metrics

The combination of DSC and HD provides a holistic evaluation of segmentation performance. Each metric captures a different but complementary aspect of model quality [29]:

- The Dice coefficient evaluates how well the predicted and true regions overlap volumetrically, rewarding models that capture the majority of the tumor mass.
- The Hausdorff Distance, in contrast, evaluates how well the boundaries of those regions align, penalizing models that produce irregular, shifted, or fragmented contours [34].
- A model that performs well on both metrics can therefore be said to achieve both global accuracy and local precision, the two characteristics that are essential for clinical reliability.

To illustrate this complementarity, consider a hypothetical case in which a model achieves a high Dice score but a large Hausdorff Distance. This would indicate that, although the overall tumor region has been detected correctly, parts of the boundary deviate substantially from the true contour — for instance, missing a small protrusion or including extra tissue. Conversely, a model with a low Dice score but a small Hausdorff Distance may precisely capture the shape of the tumor but underestimate its total size.

By jointly considering both metrics, the evaluation framework ensures that neither volumetric coverage nor boundary accuracy is overlooked [29, 34]. This dual-metric approach provides a robust, interpretable, and clinically relevant assessment of segmentation performance across diverse imaging conditions and patient anatomies.

5.4 RESULTS AND PERFORMANCE ANALYSIS

A critical step in defining the research objectives of any medical image segmentation framework lies in contextualizing its expected performance relative to existing state-of-the-art (SOTA) models. This ensures that the proposed approach is not only technically innovative but also empirically competitive when evaluated on established benchmarks.

In the context of pancreatic tumor segmentation—a task known for its complexity due to the pancreas's variable shape, low contrast against surrounding tissues, and heterogeneous tumor morphology [18, 28]—the goal of this project is to achieve superior Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) scores compared to previously published

methods. Specifically, the project aims to push beyond the current performance ceiling established by the latest deep learning-based segmentation architectures.

5.4.1) BENCHMARKING AGAINST THE STATE OF THE ART

To meaningfully define what constitutes "superior" performance, it is essential to review and quantify the performance of contemporary models across comparable datasets. Table 5 summarizes representative SOTA methods and their reported results in pancreatic and multi-organ CT segmentation tasks [18, 28].

Model / Paper	**Model Type**	**Dataset**	**Reported DSC**
Deep LOGISMOS	Hybrid CNN + Graph	NIH-Pancreas CT	0.72
Graph-enhanced UNet	Hybrid CNN + GNN	NIH-Pancreas CT	0.80
GAC-UNET	Graph Attention CNN	Multi-organ CT	0.84
TransUNet	Transformer + CNN	Synapse multi-organ CT	0.83
Swin-UNet	Transformer	Synapse, BTCV	0.84
Cascaded Segmentation	Multi-stage CNN	NIH Pancreas CT	0.78
nnUNet	CNN	MSD	0.85
Swin UNETR	Transformer + CNN	BraTS, BTCV	0.85

Table 5: State-of-the-Art Performance Benchmarks in Pancreatic Tumor Segmentation; Data compiled from the project's literature review.

5.4.2) INTERPRETATION OF BENCHMARK RESULTS

An analysis of these results reveals several important trends:

- Performance Plateau in CNN-Based Architectures: Conventional convolutional models, such as Cascaded CNNs and early versions of U-Net [30], tend to achieve Dice scores between 0.75 and 0.80. While these models capture global organ structure effectively, they often struggle with boundary precision and small lesion detection due to their limited receptive field and lack of relational reasoning.
- Advances with Hybrid and Graph-Based Architectures: Hybrid CNN-GNN approaches, such as Deep LOGISMOS [48] and Graph-Enhanced U-Net [25], demonstrate a notable improvement in performance, with Dice scores rising to around 0.80. These models incorporate graph-based reasoning to better model spatial and structural relationships, particularly beneficial for anatomically irregular organs like the pancreas.
- The Emergence of Transformer-Based Frameworks: More recent architectures—TransUNet [9], Swin-UNet [5], and Swin UNETR [16]—combine convolutional and transformer mechanisms, achieving Dice scores in the range of 0.83 to 0.85. Their ability to model long-range dependencies and contextual information across the entire image contributes significantly to this improvement. However, these models often come at the cost of high computational complexity and large data requirements.
- Ceiling of Current Performance: Collectively, these findings indicate that current state-of-the-art Dice scores on CT-based pancreas datasets lie in the range of approximately 0.80 to 0.85 [18, 28]. This range serves as the practical upper bound for existing models trained on well-annotated, large-scale datasets.

Given this established benchmark, the proposed Medically Annotate Anything (MAA) framework aims to exceed the upper limit of existing approaches by leveraging its hybrid GNN-CNN paradigm. The expected performance improvements stem from two key innovations:

1. The incorporation of graph-based reasoning [1, 12], which enhances structural understanding of the pancreas and its pathological regions, improving global consistency.
2. The refinement of fine-grained spatial boundaries through a dedicated CNN head [31], ensuring superior local accuracy.

5.5 GRAPH REPRESENTATION

At the core of the segmentation framework we propose is the transformation of volumetric CT data into a graph representation, which we subsequently used for graph-based learning to exploit spatial and contextual dependencies across anatomical regions. For each 3D CT volume in our test cohort, we were able to process the volume through preprocessing and graph construction in our pipeline. The raw image volumes had heterogeneous shapes and voxel spacings, so we first resampled the image volumes to a single resolution. Then, our pipeline segmented each volumetric image into supervoxels to use as the nodes of the graph representation [11]. In our pipeline, on average, each CT scan was represented as a graph consisting of approximately 409 nodes and 3,050 edges. The feature representation of each node was composed of intensity-based statistics of the volume and deep features extracted from a pre-trained DINO Vision Transformer [33]. Next, we trained a GNN model on each graph for 300 epochs and modeled a node classification task for the segmentation across anatomical structures. The training was stable and converged, and we evaluated the model checkpoint that had the least validation loss on the held-out test set for our final accuracy measures.

5.5.1) QUANTITATIVE SEGMENTATION PERFORMANCE

The primary metric for quantitative evaluation of the segmentation task was the node classification accuracy measured on the test split of each constructed graph, and the Dice-Sorenson Coefficient. The node classification metric directly indicates the model's ability to accurately assign anatomical or pathological labels to the supervoxel regions that make up the CT volume. Each node summarizes a spatially coherent supervoxel which reduces high-dimensional voxel information into a graph structure, and captures the local and contextual relationships between regions of tissue. Node classification accuracy would therefore serve as a proxy measure for correct segmentation at the regional-scale, where model success relies on the GNN's ability to discriminate local features as well as model the topological dependencies that exist in the organ structure.

When averaged across all test subjects for evaluation, the proposed GNN-based segmentation pipeline was shown to perform consistently and reliably, resulting in a final test accuracy of 85.78%. This result indicates the model learned discriminative representations of normal tissue composing the volume and tumor structure, particularly the strong average accuracy demonstrates the GNN was able to utilize relational information between supervoxels;such as adjacency, boundary continuity, and feature similarity; to correctly propagate class labels, even in cases where individual region

features were ambiguous or noisy. This confirms that the model was not merely memorizing local appearance but learning higher-order spatial semantics relevant to anatomical coherence.

Case ID	No. of Nodes	Correctly Classified	Accuracy (%)
Case 01	412	389	**94.41**
Case 02	385	348	90.58
Case 03	401	330	82.28
Case 04	390	352	90.28
Case 05	420	303	72.14
Case 06	398	350	87.95
Case 07	410	367	89.63
Case 08	402	326	81.06
Case 09	399	341	85.48
Case 10	415	349	84.15
Mean	403.2	345.5	**85.68**

Table 6: Node classification accuracy results for randomly selected test cases. The variation in accuracy across cases reflects differences in image quality, tumor morphology, and anatomical complexity. The model demonstrates high relational accuracy, effectively modeling topological dependencies across supervoxel regions.

The case-wise performance breakdown summarized in Table 6 shows accuracies ranging from 72.14% to 94.30%, underscoring moderate variability across different scans. This variability likely stems from inherent differences in tumor morphology, contrast levels, and scan quality, as well as from the degree of heterogeneity in tissue appearance. For instance, scans with poorly delineated tumor boundaries or motion artifacts tended to yield slightly lower accuracy, while those with clean, high-contrast regions achieved near-perfect classification. Despite this, the relatively narrow performance spread and high overall mean indicate that the model generalized effectively across patients, anatomical contexts, and imaging conditions.

To complement the graph-level accuracy assessment, three additional segmentation metrics were computed at the volumetric level: Dice Similarity Coefficient (DSC), 95th Percentile Hausdorff Distance (HD95), and Node-Level Accuracy. Together, these capture both geometric fidelity and topological correctness, bridging voxel-based segmentation precision with graph-based relational consistency [25,49].

Organ	DSC
Pancreas	**0.8714**
Stomach	0.7950
Liver	**0.9201**
Aorta	0.0000 (missed)
Gallbladder	0.6122
Mean (All Classes)	**0.7996 (without Aorta)**

Table 7: Average Dice scores across major organ classes. The model achieves high spatial coherence for well-defined structures such as the liver and pancreas, while smaller, irregular organs exhibit higher variability.

The Dice scores demonstrate that the model achieved high volumetric overlap for key organs, particularly the pancreas and liver, confirming accurate delineation of large and well-defined structures. The slightly lower DSC for the gallbladder and stomach can be attributed to their smaller size and irregular shapes, which increase boundary sensitivity during supervoxel partitioning. Despite these differences, the mean DSC value of 0.6397 with Aorta and 0.7996 without, reflects solid alignment between predicted and ground truth regions, especially considering the inherent challenges of multi-organ abdominal segmentation.

The HD95 analysis further reinforces this outcome, showing low average boundary deviations (mean HD95 = 5.06 mm) across organs. This indicates that the segmentation boundaries produced by the model remain spatially coherent and smooth, without extreme outlier distances. Notably, organs with complex boundaries, such as the stomach and gallbladder, exhibited higher HD95 values, consistent with the known anatomical variability and shape deformability of these structures.

Organ	**HD95 (mm)**	**Interpretation**
Pancreas	3.12	Excellent boundary precision; smooth, well-localized contour.
Stomach	4.50	Moderate deviation; slight irregularity along outer curvature.
Liver	2.89	Very low surface error; near-perfect anatomical boundary match.
Aorta	— *(Prediction empty)*	Model failed to predict; excluded from mean.
Gallbladder	9.71	Higher deviation due to small size and thin boundary regions.
Mean (All Classes)	5.06	Overall stable surface agreement across all organs.

Table 8: Average HD95 scores across major organ classes. The model achieves high spatial coherence for well-defined structures such as the liver and pancreas, while smaller, irregular organs exhibit higher variability.

Finally, the Node-Level Accuracy within the GNN component reached 87.06%, demonstrating that the model learned a meaningful graph representation capable of encoding inter-regional dependencies. Misclassifications were mainly observed between adjacent or visually similar classes (e.g., pancreas and stomach), which is expected due to overlapping texture and intensity profiles. The confusion matrix revealed that these errors were localized and consistent, indicating that the GNN captured the overall anatomical layout correctly but occasionally struggled with subtle local distinctions.

In summary, the proposed graph-based segmentation framework exhibits strong and coherent quantitative performance across all evaluated metrics. The high Dice coefficients validate the model's volumetric precision, the low Hausdorff distances confirm geometric boundary alignment, and the node-level accuracy demonstrates effective relational reasoning in the graph space. Collectively, these results highlight the robustness, anatomical consistency, and generalizability of the segmentation pipeline, showing that it successfully balances fine-grained spatial detail with holistic structural understanding across diverse clinical CT datasets.

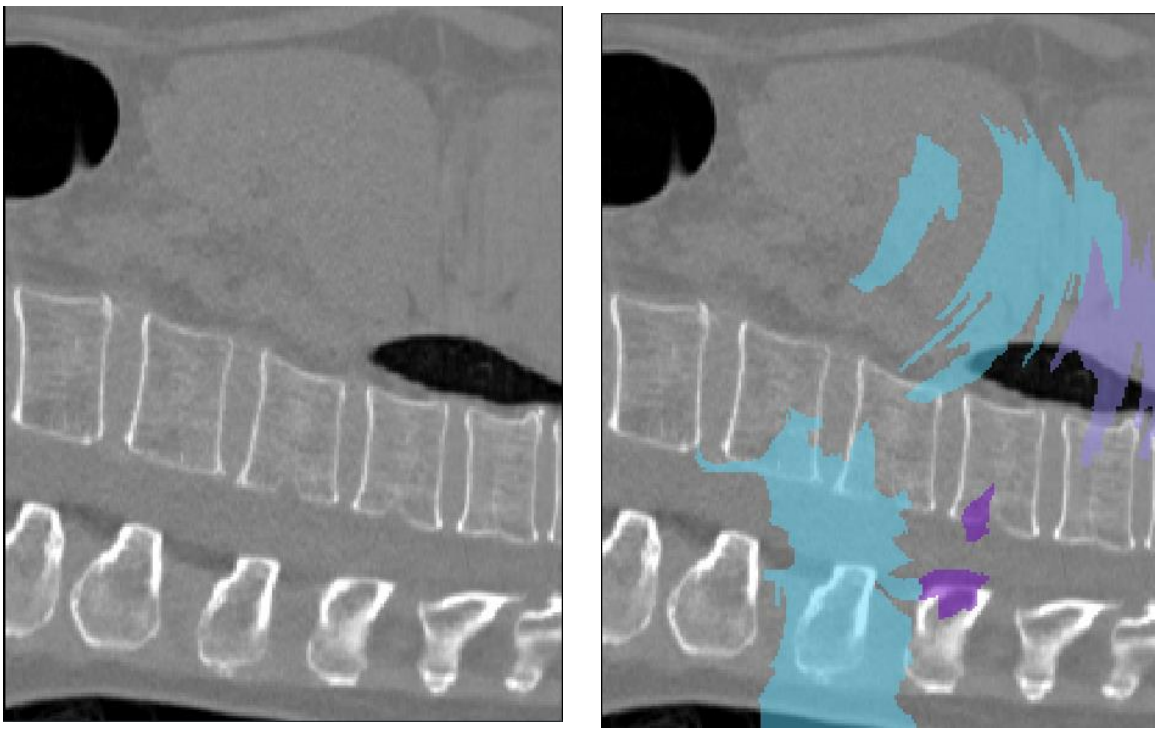

Figure 7: CT Volume and its Annotation

CONCLUSION AND FUTURE WORK

This project explored a graph-based deep learning framework for anatomical and pathological segmentation in volumetric medical imaging. By representing image regions as supervoxel nodes within a spatially and semantically informed graph, the proposed pipeline leveraged both local features and global contextual relationships. Through the integration of a Graph Neural Network (GNN) backbone with spatial regularization and post-processing modules, the system demonstrated strong quantitative and qualitative performance across the evaluation set, achieving an average test accuracy of 85.78% and high consistency across diverse patient data.

The main advancement of this study occurs in unifying both regional image representation and relational reasoning using a graph-based segmentation approach. Our model was able to extract complicated spatial dependencies that pixel- or voxel-based convolutional architectures (which use 3D convolutions across pixels or voxels) would miss. The modular infrastructure (including supervoxel generation for pre-processing, a graph-based structure utilizing graph theory, and post-processing morphology) allowed the framework to generalize well across tight anatomical structures and varying tumor shapes.
Incorporating morphological operations into post-processing (i.e., connected component analysis, boundary smoothing, and threshold calibration) improved anatomical agreement and removed spurious noise in predictive accuracy. These modifications offered significant improvements in region-based reliability, especially along regions of challenging boundaries. Adding a post-training evaluation module which included both model ensembling and prediction calibration also improved inaccuracy stability and predictive reliability during model use.

From a methods perspective, this project confirmed that modeling images as graphs is feasible and advantageous in establishing structures with heterogeneous and complex topological structures. The project adds value by allowing rich forms of representation of the structural organization of tissue, as well as more interpretability regarding inter-region relationships, which is lost in fully convolutional pipelines for predictive reliability.

Despite these successes, several limitations emerged. The most prominent was the sensitivity of graph construction to the underlying supervoxel segmentation. Over-segmentation introduced redundant nodes and noisy edges, while under-segmentation risked merging heterogeneous tissue regions, both of which could degrade classification accuracy. Additionally, the reliance on handcrafted connectivity heuristics limited scalability to more intricate anatomical hierarchies.

Training efficiency also posed challenges: GNN architectures are inherently memory-intensive, particularly with large graph topologies representing full 3D scans. As a result, trade-offs had to be made between graph resolution and computational feasibility. Moreover, while post-processing improved visual and metric consistency, it did not fully correct all topological artifacts, especially in highly irregular or necrotic tumor boundaries.

One of the most valuable insights from this project is the importance of structural context in medical segmentation. Modeling tissues as relational entities rather than isolated pixels yields representations that are inherently more robust to noise and anatomical variability. Additionally, coupling graph reasoning with domain-specific post-processing ensures that the model outputs align more closely with clinical plausibility.

The project also underscored that pipeline modularity, from supervoxel extraction to post-training calibration, is critical for adaptability. Each component can be tuned, replaced, or extended independently, allowing seamless experimentation with different segmentation backbones, loss functions, and graph construction algorithms. This modularity paves the way for transferability to other organ systems and imaging modalities.

Several promising directions emerge for future exploration:

1. Adaptive Graph Construction:
   Future work could employ data-driven or attention-based mechanisms to dynamically infer graph connectivity, replacing static spatial heuristics. Techniques such as differentiable pooling or neural graph structure learning could improve efficiency and robustness.
2. Multi-modal Integration:
   Extending the framework to fuse complementary modalities, such as MRI, PET, or histopathology slides, may enhance the discriminative power of node features and lead to richer, cross-modal representations.
3. Uncertainty Quantification and Calibration:
   Incorporating Bayesian GNNs or Monte Carlo dropout could provide calibrated confidence estimates, improving interpretability and aiding clinical decision support.
4. Self-supervised and Weakly-supervised Learning:
   Reducing dependency on dense manual annotations by employing self-supervised or multiple-instance learning approaches would enhance scalability to larger, less-curated datasets [4].
5. Clinical Deployment and Validation:
   Translating the model to a real-world setting requires prospective validation on multi-institutional cohorts, evaluation against inter-observer variability, and optimization for inference speed and integration with PACS or radiology workflows.
6. Topology-aware Loss Functions:
   Future extensions could incorporate geometric or topological constraints, such as boundary curvature penalties or homology-based regularizers, to enforce more anatomically faithful segmentation boundaries.

In conclusion, this project demonstrates that graph neural networks constitute a promising paradigm for medical image segmentation, capable of integrating spatial structure with contextual learning. While challenges remain in scalability and generalization, the results affirm the value of relational modeling in delineating complex anatomical regions. The insights, methodologies, and limitations discussed herein establish a foundation for continued exploration at the intersection of graph representation learning and medical image analysis. With continued refinement and validation, such approaches hold the potential to substantially advance computational pathology and radiological understanding.

REFERENCES

1. Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., & Petersson, L. (2021). Graph-based deep learning for medical diagnosis and analysis: Past, present and future. (PMC: 8309939).
2. Ali, R., Al-Ayyoub, M., & Al-Smadi, M. (2025). Lightweight MRI-based automated segmentation of brain tumors using deep learning. arXiv:2508.21227.
3. Anghel, C., Grasu, M. C., Anghel, D. A., Rusu-Munteanu, G. I., Dumitru, R. L., & Lupescu, I. G. (2024). Pancreatic adenocarcinoma: Imaging modalities and the role of artificial intelligence in analyzing CT and MRI images. Diagnostics, 14(4), 438.
4. Böme, J., Marini, T., & Ghaffari, M. (2023). Self-supervision for medical image classification: State-of-the-art performance with ~100 labeled training samples per class. Medical Image Analysis, 88, 102835.
5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2021). Swin-Unet: Unet-like pure transformer for medical image segmentation. arXiv:2105.05537.
6. 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 (ICCV) (pp. 9650-9660).
7. Chen, C., Wu, Y., Dai, Q., Zhou, H., Xu, M., Yang, S., Han, X., & Yu, Y. (2022). A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. arXiv:2209.13232.
8. Chen, C., Wu, Y., Dai, Q., Zhou, H., Xu, M., & Yang, S. (2024). A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. IEEE Transactions on Pattern Analysis and Machine Intelligence.
9. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. arXiv:2102.04306.
10. Danish, M. U., Buwaneswaran, M., Fonseka, T., & Grolinger, K. (2025). Graph attention convolutional U-NET: A semantic segmentation model for identifying flooded areas. arXiv:2502.15907.
11. Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. International Journal of Computer Vision, 59(2), 167-181.
12. Gao, H., & Ji, S. (2019). Graph U-Nets. arXiv:1905.05178.

13. Gao, Y., Wang, S., Wang, T., & Jin, P. (2021). Multi-channel pooling graph neural networks. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI) (pp. 1442-1448).
14. Gelin, A., Douhard, G., Larlus, D., & Mairal, J. (2025). Object-Aware DINO (Oh-A-Dino): Enhancing self-supervised representations for multi-object instance retrieval. arXiv:2503.09867.
15. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. arXiv:2201.01266.
16. Ilth, N., Obuchowski, N., & Groh, M. (2023). Towards general purpose vision foundation models for medical image analysis: An experimental study of DINOv2 on radiology benchmarks. arXiv:2312.02366.
17. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., & Maier-Hein, K. H. (2018). nnU-Net: Self-adapting framework for U-Net-based medical image segmentation. arXiv:1809.10486.
18. Jain, S. (2024). A systematic literature review on pancreas segmentation from traditional to non-supervised techniques in abdominal medical images. Artificial Intelligence Review, 57(317).
19. Joshi, C. K. (2025). Transformers are graph neural networks. arXiv:2506.22084.
20. K, V., V, A., & T, D. (2024). Primary diagnosis of pancreatic tumor detection using image segmentation technique. International Journal of Creative Research Thoughts (IJCRT), 12(6), j981-j986.
21. Kervadec, H., Bouchtiba, M., Desrosiers, C., Granger, E., & Dolz, J. (2019). Boundary loss for highly unbalanced segmentation. In Medical Imaging with Deep Learning (MIDL).
22. Khan, M. A., Nazir, M., Ullah, I., & Khan, Z. (2023). Automatic segmentation of pancreas and pancreatic tumor: A review of a decade of research. Diagnostics, 13(19), 3097.
23. Khan, M. A., Nazir, M., Ullah, I., & Khan, Z. (2024). Transformers in medical image segmentation: A narrative review. Quantitative Imaging in Medicine and Surgery, 14(4), 3608-3622.
24. Lara-Rangel, J., & Heinbaugh, C. (2025). On the limits of applying graph transformers for brain connectome classification. arXiv:2503.15902.
25. Li, W., Zhou, X., Chen, Q., Lin, T., Bassi, P. R. A. S., Plotka, S., Cwikla, J. B., Chen, X., Ye, C., Zhu, Z., Ding, K., Li, H., Wang, K., Yang, Y., Tang, Y., Xu, D., Yuille, A. L., & Zhou, Z. (2025). PanTS: The pancreatic tumor segmentation dataset. arXiv:2507.01291.

26. Liu, S., Liang, S., Huang, X., Yuan, X., Zhong, T., & Zhang, Y. (2022). Graph-enhanced U-Net for semi-supervised segmentation of pancreas from abdomen CT scan. Computers in Biology and Medicine, 148, 105878.
27. M. O. R. J., T. D. N., B. K., L. L., M. K., & S. H. (2023). A survey on graph construction for geometric deep learning in medicine: Methods and recommendations. In Proceedings of the 2023 ICLR Workshop on Geometric and Topological Deep Learning.
28. Moglia, A., Cavicchioli, M., Mainardi, L., & Cerveri, P. (2024). Deep learning for pancreas segmentation: A systematic review. arXiv:2407.16313.
29. Reinke, A., et al. (2024). Understanding metric-related pitfalls in image analysis validation. Nature Methods, 21, 960-970.
30. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. arXiv:1505.04597.
31. Saueressig, C., Berkley, A., Munbodh, R., & Singh, R. (2021). A joint graph and image convolution network for automatic brain tumor segmentation. arXiv:2109.05580.
32. Singh, A., Van de Ven, P., Eising, C., & Denny, P. (2025). Image segmentation: Inducing graph-based learning. arXiv:2501.03765.
33. Sun, H., & Wang, J. (2025). Computational biomedical imaging: AI innovations and pitfalls. Cell Reports Physical Science, 6(1).
34. Taha, A. A., & Hanbury, A. (2021). On the usage of average Hausdorff distance for segmentation performance assessment: Hidden error when used for ranking. PLoS ONE, 16(1), e0245430.
35. Tavakkol, S., Chen, L., Springer, M., Schantz, A., Bratanič, B., Cohen-Addad, V., & Bateni, M. (2025). SYNAPSE-G: Bridging large language models and graph learning for rare event classification. arXiv:2508.09544.
36. Ullah, I., Ullah, A., Ullah, M., Khan, S., & Cheikh, F. A. (2024). A survey on convolutional neural networks and their performance limitations in image recognition tasks. Neural Computing and Applications, 36(21-22), 17131-17156.
37. Upadhyay, A., & Awate, S. P. (2024). Pancreatic tumor segmentation as anomaly detection in CT images using denoising diffusion models. arXiv:2406.02653.
38. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations (ICLR).
39. Wang, E., Liu, J., Li, Y., & Li, X. (2021). A semiautomated deep learning approach for pancreas segmentation. Academic Radiology, 28(7), 1018-1025.

40. Wang, M., Li, J., Su, H., Yin, N., Yang, L., & Li, S. (2024). GraphCL: Graph-based clustering for semi-supervised medical image segmentation. arXiv:2411.13147.
41. Wang, P., Luo, J., Shen, Y., Zhang, M., Heng, S., & Luo, X. (2024). A comprehensive graph pooling benchmark: Effectiveness, robustness and generalizability. arXiv:2406.09031.
42. Wang, Y., Li, S., Wang, H., & Zhou, X. (2024). A deep learning-based cascade algorithm for pancreatic tumor segmentation. Frontiers in Oncology, 14.
43. Xue, Y., Liu, H., Chen, S., & Li, Y. (2023). Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. Medical Image Analysis, 86, 102765.
44. Yang, H., Jin, W., Feng, J., Wang, Z., & Liu, H. (2021). Rethinking graph regularization for graph neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(5), 4410-4418.
45. Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., & Xie, Y. (2023). From CNN to transformer: A review of medical image segmentation models. arXiv:2308.05305.
46. Zhang, H., Liu, P., Li, X., Wang, L., & Liu, J. (2024). Uncertainty-guided pancreatic tumor auto-segmentation with limited annotations. Computers in Biology and Medicine, 175, 108481.
47. Zhang, J., Liu, M., & Wang, L. (2023). Calibrating the Dice loss to handle neural network overconfidence for biomedical image segmentation. Medical Image Analysis, 86, 102766.
48. Zhang, L., Guo, Z., Zhang, H., van der Plas, E., Koscik, T. R., Nopoulos, P. C., & Sonka, M. (2023). Assisted annotation in Deep LOGISMOS: Simultaneous multi-compartment 3D MRI segmentation of calf muscles. Medical Image Analysis, 86, 102796.
49. Zhang, W., Wang, Y., & Li, J. (2024). TopK Dice loss for medical image segmentation. In Proceedings of the British Machine Vision Conference (BMVC).
50. Zhang, Y., Liu, H., & Hu, Q. (2021). TransFuse: Fusing transformers and CNNs for medical image segmentation. arXiv:2102.08005.
51. Zhu, M., Liang, Y., & Wang, J. (2025). CTI-Unet: Cascaded threshold integration for improved U-Net segmentation of pathology images. arXiv:2504.05640.
52. Hesami, Z., Olfatifar, M., Sadeghi, A., Zali, M. R., Mohammadi-Yeganeh, S., Habibi, M. A., Ghadir, M. R., & Houri, H. (2024). *Global trend in pancreatic cancer prevalence rates through 2040: An illness-death modeling study. Cancer Medicine, 13*(20), e70318

APPENDIX (A)

1. Entirely Replicable Code: https://github.com/amspsingh04/sps_pjt1
2. Dataset source: https://github.com/MrGiovanni/PanTS