Graph Neural Networks for Modeling Spatial Relationships Among Anatomical Structures in Three-Dimensional Medical Scans

Zamira Palastri
Department of Computer Science, George Mason University, Fairfax, VA, USA.
zamirap@gmu.edu

Abstract

Three-dimensional medical scans contain anatomical structures whose spatial arrangement carries diagnostic and procedural information that voxel-wise segmentation networks represent only implicitly. This manuscript develops a methodological framework for modeling spatial relationships among anatomical structures in three-dimensional scans using graph neural networks. The central contribution is a geometry-aware heterogeneous graph formulation in which anatomical structures are nodes and multi-attribute spatial relations are edges encoding distance, direction, orientation, containment, adjacency, and scale. A structure-level graph is constructed from upstream segmentation candidates, and an edge-conditioned attention message-passing network propagates information along anatomically plausible relations. Relational supervision is incorporated through an edge prediction objective and a structured loss that couples node labels, landmark coordinates, and spatial consistency. The paper specifies graph construction rules, message-passing functions, attention normalization, learning objectives, inference logic, and boundary conditions, and it analyzes failure modes including missing structures, false candidates, inconsistent orientation conventions, and domain shift. The framework is intended to complement volumetric convolutional and transformer backbones by adding explicit relational reasoning at the anatomical level. The analysis indicates that structure-level graphs can provide interpretable relational evidence and can support anatomy-aware refinement, landmark inference, and quality control without requiring a dense graph over voxels.

Keywords

graph neural networks; anatomical spatial relations; three-dimensional medical imaging; geometric deep learning; relational reasoning; structure-aware segmentation

1. Introduction

Three-dimensional medical scans such as computed tomography, magnetic resonance imaging, ultrasound volumes, and intraoral surface scans encode anatomy as a spatially organized collection of structures. Clinical interpretation depends not only on the appearance of individual organs, vessels, lesions, or landmarks but also on their relative position, orientation, containment, adjacency, and scale. Most contemporary segmentation systems process these scans as dense voxel grids and learn local appearance patterns through convolutional kernels or transformer blocks. Such models can achieve high voxel overlap, yet they do not explicitly represent the relational structure that distinguishes an anatomically plausible configuration from a locally plausible but globally inconsistent one.

Graph neural networks provide a natural language for relational data. Graph convolutional networks [1], attention-based graph networks [2], and neural message passing

[3] learn node representations by exchanging information along edges, and the broader graph-network formalism [4] makes relational inductive biases explicit. In three-dimensional medical imaging, structure-level relations are often implicit in the output of segmentation or landmark-detection pipelines; recent work on segmentation-guided semi-supervised landmark detection on intraoral three-dimensional scans [5] illustrates how structural cues can be coupled with landmark inference. Geometric deep learning [6] and medical image analysis surveys [7] further motivate representations that respect non-Euclidean structure and anatomical priors. Volumetric backbones such as 3D U-Net [8], nnU-Net [9], and Swin UNETR [10] supply strong candidate masks and features, but they generally do not enforce explicit relations among the structures they segment.

This paper addresses that gap by formulating a geometry-aware heterogeneous graph neural network for structure-level spatial reasoning in three-dimensional medical scans. The central contribution is not a new segmentation backbone but a relational representation and inference framework: anatomical structures become nodes, spatial relations become typed edges with geometric attributes, and message passing is conditioned on those attributes. The framework specifies how to construct the graph from scan-derived candidates, how to encode spatial relations without discarding patient orientation or physical scale, how to propagate information with edge-conditioned attention, and how to train and infer structure labels and landmarks under relational consistency. The manuscript is analytical and methodological; it does not report a clinical trial, benchmark evaluation, or fitted model. Instead, it develops the design logic, equations, assumptions, and failure modes that determine whether such a graph model is technically defensible.

2. Problem Formulation and Anatomical Graph Construction

Let I denote a three-dimensional medical scan with voxel spacing s = (s_x, s_y, s_z). An upstream segmentation, detection, or atlas-based pipeline produces a set of N anatomical candidates S = {S_1, ..., S_N}. Each candidate may be an organ mask, a vessel segment, a lesion candidate, a landmark neighborhood, or a region proposed by a registration prior. For each candidate S_i, a node representation x_i is formed from three sources: geometric descriptors such as centroid c_i, volume V_i, principal-axis orientation θ_i, bounding-box extent, and surface-area-to-volume ratio; appearance descriptors pooled from a volumetric convolutional or transformer feature map; and optional anatomical priors such as structure identity probabilities from a population atlas. The graph is defined as follows.

Here V is the node set, E is the edge set, X ∈ R^{N×d_x} stacks node features, and Z ∈ R^{|E|×d_z} stacks edge features. The graph is heterogeneous: nodes may carry types such as organ, vessel, lesion, or landmark, and edges may carry types such as containment, adjacency, distance, or directional relation. A directed edge (i, j) indicates that the spatial relation of j relative to i is available to node i; bidirectional edges are used when the relation is informative in both directions.

Edge features are the principal mechanism by which the graph represents spatial relationships. For an ordered pair (i, j), a compact edge descriptor is constructed from normalized distance, direction, orientation agreement, volumetric overlap, adjacency, and scale ratio.

$$z_ij = \left[d_ij, u_ij, \cos\theta_ij, o_ij, a_ij, \log\left(\frac{V_j}{V_i}\right) \right]$$

In this expression, d_ij = ‖c_j − c_i‖_2 / s̄ is the centroid distance normalized by a characteristic spacing s̄ derived from the scan, u_ij = (c_j − c_i) / ‖c_j − c_i‖_2 is the unit direction in patient coordinates, cosθ_ij is the cosine between dominant principal axes, o_ij is the intersection-over-union or containment ratio between candidate masks, a_ij is a binary or soft adjacency indicator based on surface distance, and log(V_j/V_i) captures relative scale. The direction term preserves patient orientation, which is important because left-right, anterior-posterior, and superior-inferior relations are not interchangeable. The scale term prevents large organs from dominating message passing simply because their masks contain more voxels.

Graph construction should be anatomy-aware rather than purely k-nearest-neighbor. A practical rule set combines geometric proximity, mask adjacency, and prior anatomical relations: for example, edges can be added between structures whose normalized centroid distance is below a threshold, between structures whose masks are within a surface-distance tolerance, and between structures that are expected to be adjacent in a population atlas. A null-node mechanism can represent missing structures so that the absence of an expected relation is encoded rather than silently ignored. This design choice matters because a missing gallbladder, an unlabeled vessel segment, or a false lesion candidate changes the graph topology, and the model must be able to distinguish absence from irrelevance.

Compared with point-cloud methods that operate on surface samples [11] or dynamic graph construction over point neighborhoods [12], the proposed graph operates at the anatomical-structure level. Point-based models are appropriate for shape analysis, but anatomical structures are volumetric and semantically defined; a structure-level graph reduces the number of nodes from millions of voxels or points to tens or hundreds of structures. General graph neural network surveys [13] and reviews [14] describe message passing and pooling mechanisms that transfer directly to this setting, while medical imaging reviews [15] emphasize the need for domain-specific inductive biases when applying such models to clinical data.

3. Geometry-Aware Graph Neural Network Architecture

The architecture uses an edge-conditioned attention message-passing scheme. Each node i has a hidden state h_i^(l) at layer l, initialized from x_i. At each layer, neighboring nodes send messages that depend on both the sender state and the edge feature z_ij. The node update is written as follows.

$$h_i\hat{}(l+1) \;=\; \varphi\hat{}(l)\big(\, h_i\hat{}(l), \sum_\{j \in N(i)\}\, \alpha_ij\hat{}(l)\, \psi\hat{}(l)(h_j\hat{}(l), z_ij)\,\big)$$

Here N(i) is the set of neighbors of i, ψ^(l) is a message function, φ^(l) is an update function, and α_ij^(l) is a normalized attention weight. Both ψ^(l) and φ^(l) can be implemented as multilayer perceptrons. The edge feature z_ij is injected into the message function so that the same sender state can produce different messages depending on whether the relation is containment, adjacency, or long-range directional alignment. In a heterogeneous graph, ψ^(l) can be instantiated separately for each edge type, or it can share parameters across types while receiving a type embedding.

Attention weights are computed from the sender state, receiver state, and edge feature, which allows the model to down-weight noisy or anatomically implausible relations.

$$\alpha_i j = \exp \frac{\left(a^T \sigma\left(W\left[h_i || h_j || z_i j\right]\right)\right)}{\sum_{k \in N(i)} \exp\left(a^T \sigma(W[h_i || h_k || z_i k])\right)}$$

In this equation, W is a learnable projection matrix, a is a learnable attention vector, σ is a nonlinear activation such as LeakyReLU, and || denotes concatenation. The softmax is taken over the neighborhood of i. This formulation is a direct adaptation of graph attention [2] with edge features, which are essential in the anatomical setting because two structures with similar appearance may have very different spatial relations. The attention weights also provide a form of interpretability: a clinician or engineer can inspect which neighboring structures contributed most strongly to a predicted label or landmark.

After L message-passing layers, node representations h_i^(L) encode both local appearance and relational context. Three output heads are natural. A node-classification head predicts the anatomical label of each candidate. An edge-relation head predicts or verifies relation labels such as containment, adjacency, or directional order. A landmark-regression head predicts coordinates or coordinate offsets for structure-specific landmarks. These heads can be trained jointly. The architecture can be attached to upstream volumetric segmentation modules such as dense V-networks [16], spatially aggregated convolutional networks [17], or V-Net [18], which provide candidate masks and appearance features. In that integration, the graph model does not replace voxel-level segmentation; it refines structure-level identity, detects anatomically inconsistent predictions, and supplies relational evidence that a voxel-wise model cannot represent directly.

4. Learning Objectives and Relational Inference

Training requires supervision at three levels: node labels, edge relations, and task-specific outputs such as landmark coordinates. A relational prediction head maps pairs of node states to relation logits or continuous relation targets.

$$\hat{r}_ij \;=\; g(h_i\hat{}(L), h_j\hat{}(L)), L_rel \;=\; \sum_\{(i,j) \in E\}\, \ell(\hat{r}_ij, r_ij)$$

Here g is a relation function, r̂_ij is the predicted relation representation, r_ij is the target relation, and ℓ is a loss such as cross-entropy for categorical relations or smooth L1 for continuous spatial offsets. The relational loss encourages the model to preserve anatomically meaningful spatial relations even when individual node appearances are ambiguous. For example, a small low-contrast structure may be difficult to identify from intensity alone, but its predicted relation to a nearby organ can disambiguate its label.

The overall objective combines the task loss, the relational loss, and a regularizer.

$$L = L_task + \lambda L_rel + \mu L_reg$$

L_task may include node classification loss and landmark regression loss, λ and μ are nonnegative weighting coefficients, and L_reg may penalize large attention entropy, irregular graph weights, or deviations from expected spatial consistency. The weights should be selected by validation logic appropriate to the clinical task; the equation does not imply fixed values. A structured variant can replace L_reg with a spatial-consistency penalty that compares predicted inter-structure distances with their input graph distances, but such a penalty must be used cautiously because anatomically abnormal cases can legitimately violate population-level spatial priors.

Inference proceeds by applying the trained graph network to a new scan and taking the argmax of the node posterior for structure labeling.

$$\hat{y}_i \;=\; argmax_k\, p(y_i \;=\; k \mid h_i\hat{}(L))$$

Here $p(y_i = k \mid h_i^{(L)})$ is the softmax posterior over anatomical classes for node i, and $\hat{y}_i$ is the predicted label. Edge predictions can be thresholded or ranked to produce a relation graph for downstream review. Landmark coordinates, when predicted, should be accompanied by uncertainty estimates such as predictive entropy or ensembled variance. This inference rule is deliberately simple; the value of the framework lies in the representation and training objectives that produce $h_i^{(L)}$, not in a complicated decoding procedure.

It is useful to contrast this design with standard dense architectures. U-Net [19] and its volumetric variants learn hierarchical appearance features through local convolutions and skip connections. Transformer-based vision models [20], including vision transformers [21], can model long-range dependencies through self-attention, but they do so over patches or tokens rather than over named anatomical structures. The graph formulation makes the relational unit explicit: an edge exists because two structures have a spatial relation, and the edge feature states what that relation is. This explicitness is an inductive bias, not a guarantee of correctness, and it must be evaluated against the possibility that the upstream segmentation already captures enough context.

5. Analytical Assessment: Failure Modes and Design Trade-offs

The main technical risk is graph construction error. If an upstream segmentation misses a structure, the graph loses a node and all incident edges. A null-node mechanism can mitigate this by representing expected but absent structures, but it cannot recover a structure that was never proposed. False candidates create spurious nodes and can attract attention away from true structures. Dense edge construction increases recall of true relations but also admits noisy relations; sparse edge construction reduces computation but may disconnect anatomically important long-range relations such as the relative position of the heart and liver in a full-body scan. A defensible design uses a hybrid edge set: local adjacency edges, distance-threshold edges, and prior-anatomy edges, with edge-type embeddings so that the model can learn different reliabilities for each type.

A second risk is coordinate and orientation inconsistency. Medical scans vary in voxel spacing, patient position, and acquisition protocol. If direction vectors are computed in voxel coordinates without accounting for affine transforms, the model may learn scanner-specific rather than anatomical relations. Normalizing centroids by physical spacing and expressing directions in a patient-oriented coordinate frame are necessary conditions for cross-protocol use. Orientation agreement based on principal axes can also be unstable for nearly isotropic structures, so it should be treated as a soft feature rather than a hard constraint.

A third risk is label noise and relational ambiguity. Anatomical relations are not always categorical. A structure may be adjacent to another in one patient and separated by fat or bowel in another; containment may be partial; landmark relations may be defined by clinical convention rather than geometry alone. The relational loss should therefore support soft labels or ordinal targets, and the model should not be penalized for predicting a plausible alternative relation when the anatomy is genuinely ambiguous. Attention weights provide a diagnostic signal, but they are not causal explanations; high attention to a neighboring structure indicates information flow, not proof that the structure caused the prediction.

Computational cost is favorable relative to voxel-level graphs because N is typically small. The attention denominator scales with the number of edges, and a graph with tens or hundreds of nodes is inexpensive compared with dense self-attention over hundreds of thousands of voxels. The trade-off is that the graph model depends on the quality of the

upstream candidate generator. If the upstream model produces poor masks, the graph cannot compensate for missing appearance evidence. Conversely, if the upstream model is already highly accurate, the graph may add little unless the task requires relational consistency, landmark inference, or quality control.

Compared with point-cloud and mesh methods [11, 12], the structure-level graph discards fine surface detail. This is acceptable for structure labeling and relational reasoning but not for tasks that require submillimeter boundary delineation. Compared with transformer models [20, 21], the graph has a stronger prior over anatomical relations and a smaller token set, but it cannot discover relations that are not represented by edges. These are boundary conditions, not defects: they define when the framework is appropriate and when a dense voxel model or a hybrid architecture is preferable.

6. Technical Discussion and Clinical Translation

The framework is most plausible as an anatomy-aware refinement and auditing layer. A segmentation backbone produces candidate structures; the graph model predicts structure labels, verifies expected relations, and flags inconsistent configurations. In clinical review, an edge-level relation prediction can be shown alongside the image, allowing a reviewer to see which spatial relations supported or contradicted a label. This is useful in multi-organ abdominal imaging, thoracic imaging, and head-and-neck scans, where anatomy is highly structured but variable across patients. The graph schema must be defined per anatomical region; a single universal schema is unlikely to serve all tasks.

Semi-supervised and weakly supervised settings are natural extensions. If relational labels are available for a small subset of scans, the relational loss can propagate structure-level constraints to unlabeled scans through pseudo-labels and consistency regularization. The segmentation-guided semi-supervised landmark detection work on intraoral scans [5] is relevant here because it shows how structural cues and landmark inference can be coupled in a three-dimensional setting; the present framework generalizes that coupling to a graph of multiple anatomical structures. However, pseudo-labeling introduces confirmation bias, and relational consistency should not be enforced so strongly that it suppresses genuine anatomical variants.

Calibration is another translational concern. A graph model that predicts structure labels with high confidence but poor calibration is unsafe for triage. Predictive entropy, ensemble disagreement, and edge-level relation confidence should be reported alongside labels. The model should also expose which nodes were missing or assumed null, because a confident prediction built on an incomplete graph is different from one built on a complete graph. Finally, the framework should be evaluated with clinically meaningful endpoints such as structure-level precision, landmark localization error, relation accuracy, and detection of anatomically implausible outputs, not only voxel overlap. Those evaluations are future work and are outside the scope of this methodological manuscript.

7. Conclusion

This paper has developed a geometry-aware graph neural network framework for modeling spatial relationships among anatomical structures in three-dimensional medical scans. The framework represents structures as heterogeneous nodes, encodes distance, direction, orientation, containment, adjacency, and scale as edge features, and propagates information through edge-conditioned attention message passing. Relational supervision and structured losses connect node labels, landmark coordinates, and spatial consistency, while

the inference rule produces structure labels and relation predictions that can be audited. The analytical assessment identifies graph construction error, coordinate inconsistency, relational ambiguity, and dependence on upstream candidate quality as the principal boundary conditions. The contribution is a technically substantive design for explicit anatomical relational reasoning that complements, rather than replaces, volumetric segmentation and transformer backbones.

References

1. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In International Conference on Learning Representations (ICLR).

2. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations (ICLR).

3. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. In Proceedings of the 34th International Conference on Machine Learning (ICML), 1263–1272.

4. Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... & Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261.

5. Liu, S., Jin, J., Li, W., & Yan, X. (2026). Segmentation-Guided Semi-Supervised Landmark Detection on Intraoral 3D Scans. Preprints. https://doi.org/10.20944/preprints202609.1722.v1

6. Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. IEEE Signal Processing Magazine, 34(4), 18–42.

7. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), 424–432.

9. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203–211.

10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes), 272–284.

11. Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in Neural Information Processing Systems (NeurIPS), 5099–5108.

12. Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics, 38(5), 146.

13. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24.

14. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57–81.

15. Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221–248.

16. Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., ... & Barratt, D. C. (2018). Automatic multi-organ segmentation on abdominal CT with dense V-networks. IEEE Transactions on Medical Imaging, 37(8), 1822–1834.

17. Roth, H. R., Lu, L., Lay, N., Harrison, A. P., Farag, A., Sohn, A., & Summers, R. M. (2018). Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Medical Image Analysis, 45, 94–107.

18. Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the Fourth International Conference on 3D Vision (3DV), 565–571.

19. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234–241.

20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.

21. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR).