MedGraphGen: A Unified Pipeline for Generating Medical Knowledge Graphs from Medical Images

Chengzhi Cao, Min Xu*
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, United Arab Emirates
Chengzhi.Cao@mbzuai.ac.ae, Min.Xu@mbzuai.ac.ae

***Abstract*—Medical image analysis has evolved beyond single-organ segmentation towards holistic understanding of anatomical relationships. However, current approaches struggle to explicitly represent the complex structural dependencies between multiple organs, which limits their utility for downstream tasks. In this paper, we propose MedGraphGen, a unified pipeline that directly generates interpretable medical knowledge graphs from CT images in an end-to-end manner, enabling structured understanding beyond pixel-level perception. Specifically, we introduce a Spatial-Semantic Context Aggregator to bridge segmentation features with graph-based reasoning by explicitly modeling spatial relationships between organs. These spatial priors are integrated into a Relation-Aware Hypergraph Module, which leverages hyperedges to represent higher-order interactions among multiple organs. Then, we introduce Relation-Aware Hyperedge Attention (RA-HA) to differentiate semantically meaningful multi-organ interactions from trivial co-occurrences during message passing. Crucially, to support this task, we present a new benchmark dataset with dense pixel-wise segmentation masks and medical knowledge graph annotations, enabling training and evaluation of models capable of medical image understanding. We demonstrate that the generated graphs are not only anatomically plausible but also serve as powerful priors to enhance the relational reasoning capabilities of medical vision-language models.**

***Index Terms*—Knowledge Graph, Medical Image Analysis, Relational Reasoning**

I. Introduction

Computed Tomography (CT) imaging is fundamental to modern diagnosis and treatment planning, providing detailed anatomical information. A key challenge lies in moving beyond isolated organ identification toward a systems-level understanding that captures the intricate spatial and functional relationships among multiple organs [1]. Such structured anatomical knowledge is crucial not only for clinical tasks like surgical navigation but also as foundational reasoning support for advanced medical artificial intelligence, particularly Medical Vision-Language Models.

While deep learning has revolutionized organ segmentation, existing methods [2] [3] treat this as a dense per-pixel prediction problem. They excel at identifying "what" and "where" but lack an explicit mechanism to model "how" organs relate to one another. This limits their ability to provide the structured, relational context required for complex clinical reasoning. Concurrently, Medical Vision-Language Models [4], which aim to answer questions about medical images, have shown promise but often struggle with tasks requiring deep anatomical reasoning. Their performance can be hampered by a lack of explicit, grounded knowledge about organ interactions, leading to plausible but incorrect inferences.

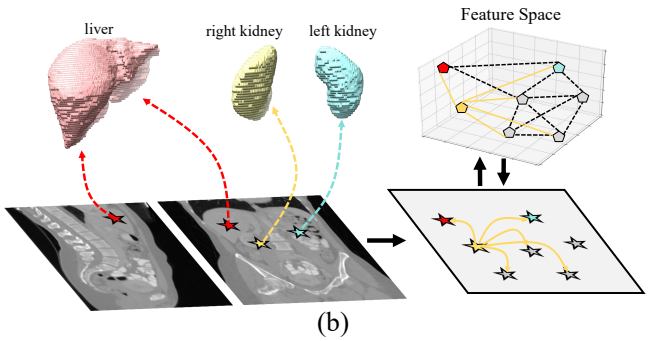


Fig. 1. We formulate the organs as nodes and their spatial correlations as edges in a structured graph representation, then employ deep learning strategies to learn both node representations and edge dependencies in the feature space.

To bridge this gap, we propose MedGraphGen, a unified end-to-end pipeline that generates interpretable medical knowledge graphs directly from CT scans, enabling structured understanding beyond isolated object detection. Our framework integrates a segmentation backbone with a novel Spatial-Semantic Context Aggregator, which extracts relative geometric cues and transforms them into initial spatial relations. These are further refined within a Relation-Aware Hypergraph Module, where hyperedges naturally represent higher-order anatomical constellations. To prioritize semantically meaningful interactions during message passing, we introduce Relation-Aware Hyperedge Attention, an attention mechanism that dynamically weights hyperedges based on their functional significance, enhancing clinically relevant multi-organ patterns. Crucially, to enable training and evaluation of such models, we contribute a new benchmark derived from Synapse [5] and TotalSegmentator [3], with both pixel-accurate segmentation masks and medical knowledge graph annotations. We demonstrate that MedGraphGen produces anatomically plausible graphs and significantly improves vision language model performance on medical VQA tasks. As shown in Fig. 1, each node is associated with an organ label and contains

*Corresponding Author

additional attributes such as shape, size, and location. The set of edges represents relationships between organs, and each edge is associated with a predicate that describes the correlation type. This formulation allows for a more structured and interpretable understanding of multi-organ dependencies, facilitating downstream medical image analysis. It enables the explicit modeling of multi-organ relationships, integrates information about organ-to-organ interactions, and offers a holistic view of organ organization.

The main contributions of our work are as follows:

- We propose MedGraphGen, a novel framework to generate interpretable medical knowledge graphs directly from CT images, bridging pixel-level segmentation with structured anatomical reasoning.
- We introduce a hypergraph-based Relation-Aware Hyperedge Attention, enabling explicit modeling of spatial relations and higher-order organ interactions through semantically weighted message passing.
- We present a new dataset with paired segmentation masks and annotated medical knowledge graphs, and demonstrate that the generated graphs serve as effective priors that significantly enhance the reasoning ability of medical vision-language models.

II. Related Work

A. Medical Image Analysis

Medical image analysis has witnessed remarkable progress in recent years, particularly in tasks such as organ segmentation [2] and anatomical landmark localization [6]. These efforts have largely focused on dense pixel-level prediction, where deep neural networks are trained to map raw imaging data to semantic labels or probability maps [7] [8]. Traditional approaches [9] utilize spatial priors from manually annotated atlases to guide segmentation, which struggle with computational efficiency and adaptability to new datasets. Fully convolutional networks [10] have demonstrated remarkable success due to their ability to capture multiscale contextual features. Despite these advances, deep learning typically rely on implicit learning of anatomical priors, which can not fully exploit the structural relationships between organs [11]. Other works [11] [12] have also integrated shape priors into neural networks, allowing more anatomically plausible predictions. In recent years, vision-language models (VLMs) have been adapted to the medical domain [13] [14]. However, they typically treat images as holistic embeddings and lack mechanisms for compositional reasoning. As a result, they often fail on queries requiring fine-grained spatial understanding.

B. Medical Graph Learning

Graph-based methods have been widely used in medical imaging to model complex relationships between anatomical structures [15] [16]. By encoding data as nodes and their relationships as edges, these methods [17] [18] leverage graph representations to encode spatial and structural information, enabling more effective feature extraction and reasoning [19]. In the context of medical images, graphs provide a natural way to model the spatial and functional relationships between organs. Recent studies [6] [8] have explored graph convolutional networks to capture inter-structure dependencies. For example, Rahman et al. [20] introduced a new graph-based cascaded convolutional attention decoder for multi-stage feature aggregation. However, they only use graphs as auxiliary constraints, rather than learning to generate rich relational structures in a data-driven manner.

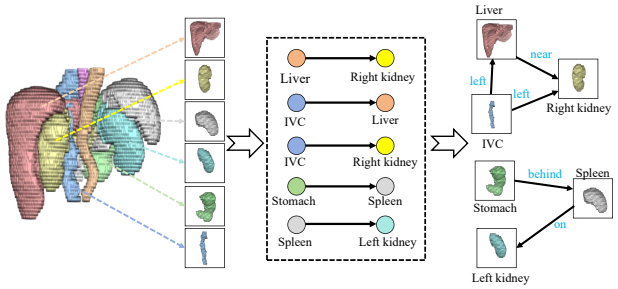


Fig. 2. Illustration of the multi-organ correlation graph. We represent organs as nodes and their inter-organ relationships as edges, capturing both spatial and contextual dependencies.

III. Method

Our proposed MedGraphGen is an end-to-end framework designed to translate medical images into structured, anatomically consistent knowledge graphs. Our approach consists of a Multi-Organ Segmenter, a Spatial-Semantic Context Aggregator, a Relation-Aware Hypergraph Module, and a Knowledge Graph Decoder. The pipeline simultaneously outputs pixel-level multi-organ segmentation maps and a structured knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, where nodes $v_i \in \mathcal{V}$ represent organ entities, and typed edges $(v_i, r_{ij}, v_j) \in \mathcal{R}$ define their explicit anatomical relationships. As shown in Fig. 2, each node represents a specific organ and anatomical structure, and each edge encodes a relationship between two organs, which can be defined based on spatial adjacency and statistical co-occurrence. Both nodes and edges are associated with attributes, such as organ shape, size, spatial proximity, and correlations. The overview of our proposed framework is shown in Fig. 3.

A. Multi-Organ Segmentation

We employ a pre-trained convolutional neural network (ResNet-50 [21]) as the visual feature extractor and generate the feature map $F_0 \in \mathbb{R}^{H \times W \times C}$. Then, it passes through the transformer-based encoder [20] with fixed positional encoding and generates the feature context $F' \in \mathbb{R}^{HW \times C}$. The feature decoder will transform organ queries, which interact with each other to capture the organ context from F', into the organ representations $X_{visual} \in \mathbb{R}^{N_o \times C}$, where N_o represents the number of organ queries. Finally, they will be fed into a segmentation head to generate high-quality segmentation predictions. The final output is the organ-level segmentation mask M and a corresponding feature embedding f_i for each segmented organ instance, extracted via masked ROI pooling

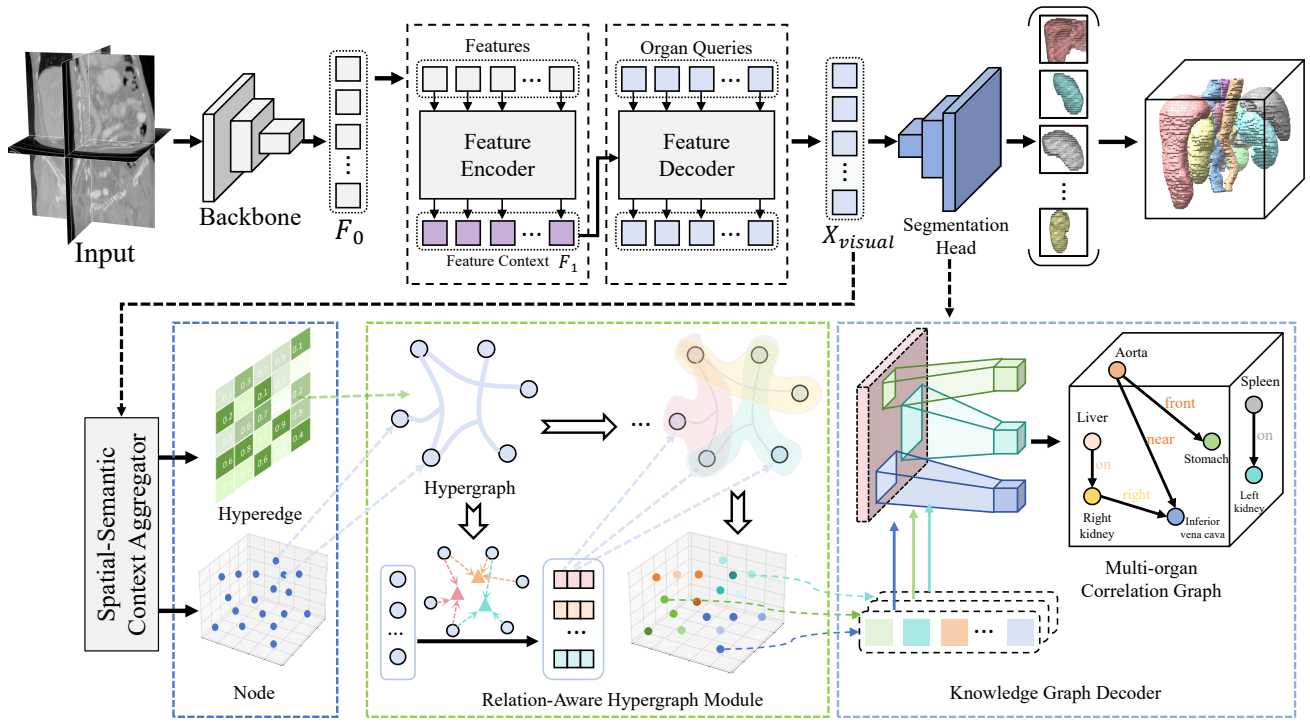


Fig. 3. The overview of the proposed architecture. Given a set of scan images, the backbone extracts multi-scale features, which are processed by a Spatial-Semantic Context Aggregator to construct hypergraph nodes and hyperedges encoding pairwise spatial and semantic relationships among organs. These are refined via a Relation-Aware Hypergraph Module to model higher-order organ interactions.

from the final feature map. Our econder is based on R50-ViT, a hybrid vision transformer model that combines a ResNet-50 [21] backbone with a Vision Transformer (ViT) [22]. The goal of this hybrid design is to leverage the feature extraction capabilities of ResNet-50 while benefiting from the long-range dependency modeling of transformers. The ViT and ResNet-50 were pretrained on ImageNet. The input resolution and patch size are set as 224×224 and 16, unless otherwise specified. The detailed structure is shown in Table I. Then we directly deploy three convolutional layers as the decoder to generate the segmentation maps.

We use cross-entropy loss for multi-class segmentation tasks, and it is defined as:

$$\mathcal{L}_{seg} = -\left(\frac{1}{I}\sum_{i}^{I} p_i \log(\hat{p}_i) + (1-p_i)\log(1-\hat{p}_i)\right), \quad (1)$$

where p_i and $\hat{p}_i$ correspond to the ground truth label for pixel i and the predicted probability for pixel i, respectively. I is the total number of pixels.

B. Spatial-Semantic Context Aggregator

This novel module is the bridge between pixel-level features and graph-level reasoning, which explicitly models the spatial context and semantic context for each organ. First, we compute the spatial context vector s_{ij} the relative geometric relationships between every pair of organs (i, j):

$$s_{ij} = [\Delta x_{norm}, \Delta y_{norm}, \Delta z_{norm}, d_{norm}], \quad (2)$$

where Δ denotes normalized centroid displacement in 3D, d_{norm} is the normalized Euclidean distance between centroids.

For semantic context vector, we initialize a learnable organ-type embedding for each of the K predefined organ classes. The semantic context for a pair (i, j) is the element-wise product of their type embeddings:

$$c_{ij} = E_{type}(class(i)) \odot E_{type}(class(j)), \quad (3)$$

where $E_{type} \in \mathbb{R}^{K\times d}$ is a learnable organ-type embedding matrix, d is dimensionality of the embedding vector (set as 128).

The initial node feature $h_i^{(0)}$ for organ i is not just its visual feature f_i, but an aggregate of its neighbors' contexts:

$$h_i^{(0)} = MLP\left(f_i || \sum_{j\in\mathcal{N}(i)} MLP(s_{ij}) || \sum_{j\in\mathcal{N}(i)} c_{ij}\right), \quad (4)$$

where $||$ denotes concatenation and $\mathcal{N}(i)$ is the set of spatially proximate organs. This step grounds each node in its local anatomical environment.

C. Relation-Aware Hypergraph Module

Many anatomical structures and functional systems are defined not by binary relations, but by coordinated interactions among groups of organs. The hypergraph generalizes edges to connect any number of nodes, providing a natural and powerful abstraction to directly model these group-wise relationships. Traditional hypergraph convolution treat all nodes

TABLE I

THE DETAILED STRUCTURE OF TH ENCODER, WHERE H AND W REPRESENT THE HEIGHT AND WIDTH OF INPUT IMAGES, N IS THE NUMEBR OF TOKENS, D IS THE EMBEDDING DIMENSION (AS 1024).

Stage	Layer Type	Number of Layers	Output Dimensions
ResNet-50	Convolution (7×7, stride=2)	1	$\frac{H}{2} \times \frac{W}{2} \times 64$
	Max Pooling (3×3, stride=2)	1	$\frac{H}{4} \times \frac{W}{4} \times 64$
	Residual Block 1 (Conv3×3)	3	$\frac{H}{4} \times \frac{W}{4} \times 256$
	Residual Block 2 (Conv3×3)	4	$\frac{H}{8} \times \frac{W}{8} \times 512$
	Residual Block 3 (Conv3×3)	6	$\frac{H}{16} \times \frac{W}{16} \times 1024$
Patch Embedding	Flatten & Linear Projection	1	$N \times D$
Transformer Encoder	Multi-Head Self-Attention	12	$N \times D$
	Feed-Forward Network (MLP)	12	$N \times D$
	Layer Normalization	1	$N \times D$
	Positional Embeddings	1	$N \times D$

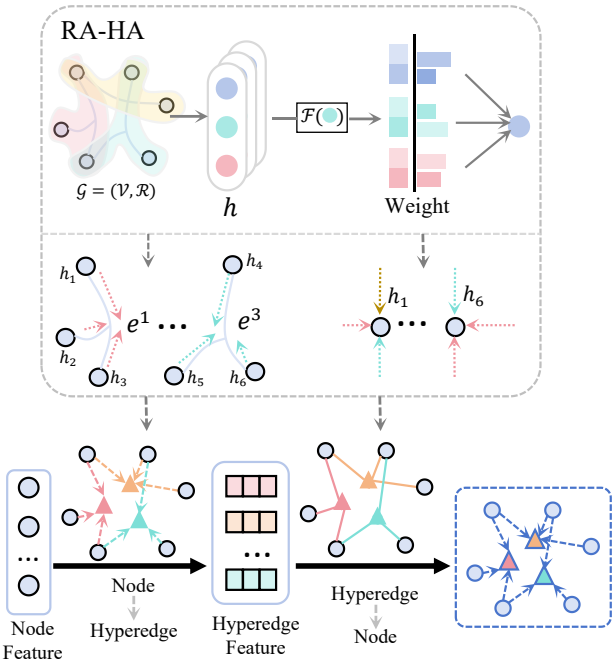


Fig. 4. The structure of Relation-Aware Hypergraph Module. "RA-HA" represents Relation-Aware Hyperedge Attention mechanism.

within a hyperedge uniformly, ignoring the rich semantic diversity of medical relationships, which are critical for accurate anatomical reasoning. To overcome this limitation, we propose a Relation-Aware Hyperedge Attention (RA-HA) mechanism (shown in Fig. 4). This module dynamically weights node contributions within each hyperedge based on both their visual features and the semantic type of the relationship encoded by that hyperedge, enabling the model to prioritize anatomically meaningful interactions during message passing. Each hyperedge corresponds to a predefined medical relation type, and we assign each relation type r_e a learnable semantic embedding $h_e \in \mathbb{R}^{d_s}$, initialized randomly and updated during training. These embeddings allow the model to distinguish between different relational semantics during aggregation. For each node $n_i \in e$, we compute an attention weight $\alpha_{i,e}$ that reflects how relevant node i is within hyperedge e, conditioned on both its own feature and the hyperedge's semantic context:

$$\alpha_{i,e} = \frac{\exp(a^T \cdot ReLU(W_q h_i + W_k h_e))}{\sum_{j \in e} \exp(a^T \cdot ReLU(W_q h_j + W_k h_e))}, \tag{5}$$

where h_i is the node feature vector, h_e is hyperedge semantic embedding, W_q, W_k are linear projections, and a is a learnable attention vector. This attention mechanism allows the model to emphasize nodes that are more semantically aligned with the hyperedge's relation type. The updated node representation at layer $l+1$ is computed as:

$$h_i^{(l+1)} = \sigma(\sum_{e \ni i} \alpha_{i,e} \cdot \sum_{j \in e} W h_j^{(l)}) \tag{6}$$

where σ is a non-linear activation as ReLU, $W \in \mathbb{R}^{d_h \times d_h}$ is a transformation matrix. After $L = 4$, we obtain contextually enriched, relation-aware node representations $h_i^{(L)}$.

D. Knowledge Graph Decoder

This module decodes the node representations into the final set of relational triplets. For every ordered node pair (i, j), we compute a score for each possible predicate:

$$\phi(i, r, j) = MLP(h_i^{(L)} || h_j^{(L)} || s_{ij} || (h_i^{(L)} \odot h_j^{(L)})). \tag{7}$$

To avoid redundant predictions and enforce structured output, we formulate graph generation as a set matching problem. Let $\mathcal{Y} = \{y_k\}_{k=1}^{T_{gt}}$ be the ground truth set of triplets, and let $\mathcal{P}_{all} = \{\phi(i, r, j)\}_{i,j=1,r \in \mathcal{P}}^{N}$ be all possible predicted triplets. We find the optimal bipartite matching $\hat{\sigma}$ between these two sets:

$$\hat{\sigma} = \arg\min_{\sigma \in G} \sum_{k=1}^{G} \mathcal{C}_{match}(y_k, \phi_{\sigma(k)}), \tag{8}$$

$$\begin{aligned} \mathcal{C}_{match}(y_k, \phi_{\sigma(k)}) = -1_{\{y_k \neq \varnothing\}} \cdot \phi_{\sigma(k)}(r_k) + \\ \lambda_{type} \cdot \mathcal{L}_{type}(s_k, o_k, \hat{s}_{\sigma(k)}, \hat{o}_{\sigma(k)}), \end{aligned} \tag{9}$$

where λ_{type} Weight for type matching loss (set as 0.01), $\mathcal{L}_{type}$ is the cross-entropy loss for matching subject/object types, s_k and o_k are the ground truth subject and object indices, respectively. $\hat{s}_{\sigma(k)}$ and $\hat{o}_{\sigma(k)}$ are predicted subject and object indices, respectively. After obtaining the optimal matching $\hat{\sigma}$, we compute the graph loss:

$$\mathcal{L}_{graph} = \sum_{k=1}^{T_{gt}} [-\log \frac{\exp(\phi(i,r,j))}{\sum_{r' \in \mathcal{P}} \exp(\phi(i,r',j))} + \lambda_{type}\mathcal{L}_{type} + \lambda_{spatial}||\hat{s}_{ij} - s_{ij}||_2^2]. \quad (10)$$

And the total loss function is computed as:

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \mathcal{L}_{graph}. \quad (11)$$

IV. Experiments

A. Datasets

Based on Synapse [5] and TotalSegmentator [3] datasets, we build two datasets for medical knowledge graph generation, called Synapse-G and TotalSeg-G dataset. Specifically, each patient's CT images are paired with high-resolution organ segmentation masks and automatically generated multi-organ knowledge graphs encoding spatial and anatomical relationships. These relational graphs are synthesized via a large language model (Qwen2.5-vl [23]) prompted with slice-specific organ lists and instructed to output clinically plausible, directional pairwise relations. Given the set of visible organs per slice, the LLM generates clinically plausible relational statements using key anatomical descriptors such as Spatial Direction ("anterior, behind, above, beneath, left_of, right_of"), Proximity ("adjacent_to, near, overlies, underlies, encapsulated_by"), Structural Relationship ("runs_parallel_to, connects_to, drains_into, supplies, innervates, is_part_of, is_located_in"). The final datasets Synapse-G has 4422 scans for training and 1715 scans for testing, and TotalSeg-G has 2040 scans for training and 380 scans for testing.

B. Metrics

In our experiments, we use Dice Score, Hausdorff Distance to assess segmentation performance. To evaluate the accuracy of knowledge graph generation, we use the F1-score as metric to ensure balanced evaluation of precision and recall across different components of the generated organ graph.

Dice Similarity Coefficient, commonly referred to as Dice Score, is used for evaluating the overlap between the predicted segmentation and the ground truth. It is defined as:

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|}, \quad (12)$$

where A and B represent the set of predicted pixels (segmentation result) and ground truth pixels, respectively. $|A \cap B|$ represents the number of overlapping pixels between the prediction and the ground truth. The Dice score of 1.0 indicates a perfect match, and 0 means no overlap at all.

Hausdorff Distance measures the worst-case discrepancy between the boundary points of the predicted segmentation and the ground truth. It is defined as:

$$\text{HD}(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\right\}, \quad (13)$$

where $d(a,b)$ is the Euclidean distance between points a and b. The function $\sup$ finds the maximum distance from a point in one set to the closest point in the other set, and the function $\inf$ finds the minimum distance from one point to another. Moreover, HD95 computes the 95th percentile of all point-wise distances, instead of taking the maximum distance. Lower HD95 values indicate better segmentation performance.

TABLE II
PERFORMANCE COMPARISON OF MULTI-ORGAN GRAPH GENERATION WITH DIFFERENT BACKBONES ON SYNAPSE-G DATASET AND TOTALSEG-G DATASET. ↑(↓) DENOTES THE HIGHER (LOWER) THE BETTER. THE BEST RESULTS ARE SHOWN IN BOLD.

Methods	Synapse-G			TotalSeg-G
	DICE↑	HD95↓	F1↑	F1↑
DARR	69.77	-	72.74	64.03
U-Net	61.50	39.61	80.26	70.41
TransUNet	77.48	31.69	88.12	72.79
AttnUNet	75.57	36.97	85.35	71.79
R50-AttnUNet	77.77	36.02	86.20	72.78
SwinUNet	79.13	21.55	87.93	72.37
MISSFormer	81.96	18.20	80.75	70.45
ViT	71.29	32.87	85.35	73.04
PVT	83.28	15.83	89.75	69.19
CASCADE	82.68	17.34	92.59	71.15
nnU-Net	80.74	15.46	90.07	72.11
Ours	**84.12**	**12.45**	**93.30**	**73.88**

TABLE III
QUANTITATIVE ANALYSIS ON MEDICAL VQA DATASETS. "*" REPRESENTS THE MODEL IS FINE-TUNED BY QUESTION-ANSWER DATA, GENERATED FROM OUR MEDICAL KNOWLEDGE GRAPHS.

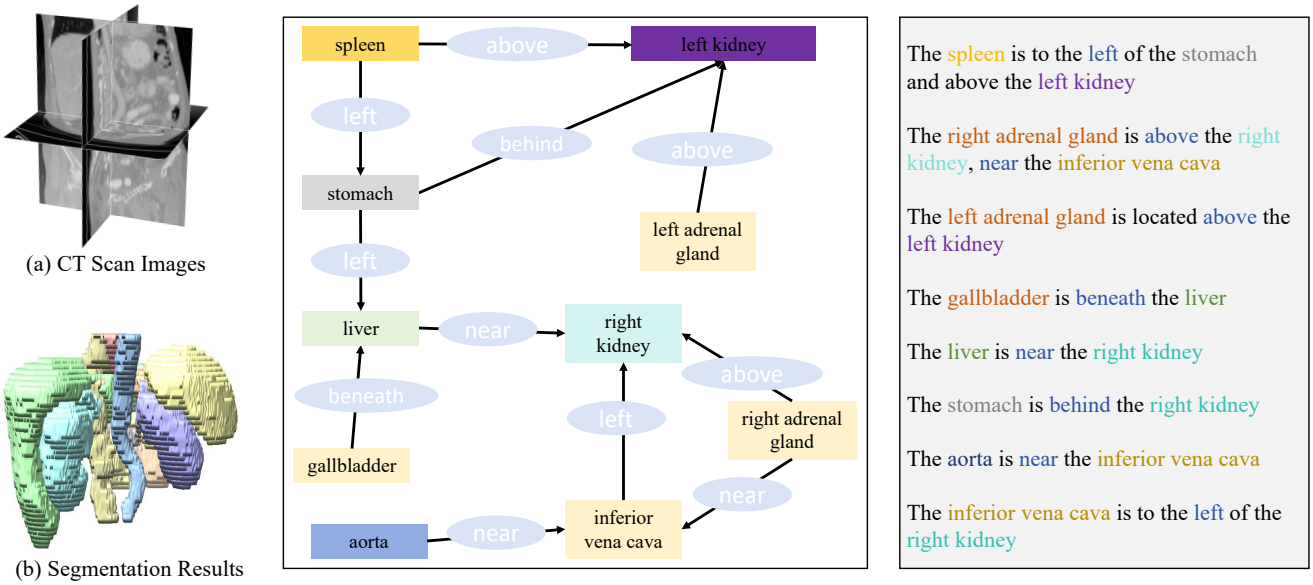
Methods	Slake	VQA-RAD
MISS	82.0	76.0
M2I2	81.2	76.8
MILE	83.0	74.8
MISS*	82.3	76.2
M2I2*	81.8	76.9
MILE*	83.5	75.1

C. Implementation Details

Our algorithm is implemented using Pytorch frameworks and trained on 4 NVIDIA GeForce RTX 3090. We optimize our algorithm based on Adam [24] optimizer, and the initial learning rate is set as 10^{-4}. We set a step scheduler to reduce the learning rate by 10 after 200 epochs, and the total number of epochs is 400. Moreover, we also compare our method with several competitive methods, including U-Net [25], DARR [26], AttnUNet [27], TransUNet [28], SwinUNet [29], MISSFormer [30], PVT [20], ViT [22], CASCADE [31], nnU-Net [32]. Because most of the compared methods are designed for segmentation, we introduce a graph head, composed of two convolutional layers and three fully connected layers to generate one-hot vectors encoding the categories of subjects, objects, and corresponding predicates. The input resolution and patch size are set as 224×224 and 16.

D. Results Analysis

Table II shows the performance of segmentation and multi-organ graph prediction. It is obvious that our method achieves the state-of-the-art performance in both datasets and outperforms all CNN-based and transformer-based methods.



(a) CT Scan Images

(b) Segmentation Results

(c) Generated Graph

(d) Description

Fig. 5. Visualization of multi-organ knowledge graph generation. (a) The input CT scan image. (b) The segmentation results of our method. (c) Qualitative results showing multi-organ knowledge graph generated by our method. (d) The explanation of the generated multi-organ knowledge graph. The relations between different organs are highlighted in blue, and organs are highlighted in different colors.

We add visualization and explanation about the generated multi-organ knowledge graph from the Synapse Dataset in Fig. 5. The prediction ⟨`spleen`, `above`, `leftkidney`⟩ and ⟨`spleen`, `left`, `stomach`⟩ mean that the spleen is to the left of the stomach and above the left kidney. Moreover, some organs have more than two relation with others, and an organ can appear as both a subject and an object simultaneously. Our framework can adapt to more complex organ systems with more predicates to analyze entire body systems.

To leverage the rich relational structure encoded in our automatically generated medical knowledge graphs, we synthesize a set of spatially grounded question-answer pairs directly from the graph edges based on SLAKE and VQA-RAD datasets. Each edge, representing a directional, proximity, or functional relationship between two organs, is mapped to natural language questions using a template-based system. For example, the triplet (Spleen, superior_to,Left Kidney) can be transferred to the question: "Is the spleen positioned superior to the left kidney?". These QA pairs are then merged with the original SLAKE and VQA-RAD training sets to form augmented datasets. As shown in Table III, fine-tuning medical VQA models (M2I2 [4], MISS [33], and MILE [34]) on the augmented datasets consistently improves performance across both benchmarks. On SLAKE, accuracy increases by 0.3–0.6% (e.g., MILE improves from 83.0% to 83.5%), while on VQA-RAD, gains range from 0.1–0.3%. Notably, even models like MISS and M2I2, which do not explicitly model spatial relations, benefit significantly, indicating that our graph-derived QA data provides valuable inductive bias for anatomical reasoning.

As illustrated in Fig. 6, when asked "From the patient's perspective, is the spleen located to the left of the liver?", the baseline model MISS incorrectly answers "No" — reflecting a common failure mode where visual position is conflated with anatomical orientation. In contrast, MISS*, trained with QA pairs synthesized from our medical knowledge graphs, correctly answers "Yes". The result confirms that augmenting VQA training with our graph-derived spatial relations enables models to overcome perspective ambiguity.

TABLE IV
COMPARISON OF GRAPH CONVOLUTIONAL NETWORK (GCN), GRAPH ATTENTION NETWORK (GAN), AND OUR HYPERGRAPH-BASED METHODS IN SYNAPSE-G AND TOTALSEG-G DATASETS.

Dataset	Synapse-G	TotalSeg-G
Graph(GCN)	85.32	70.01
Graph(GAN)	88.14	72.32
Ours	93.30	73.88

TABLE V
ABLATION STUDY OF DIFFERENT NUMBERS OF RA-HA.

Num	Synapse-G	TotalSeg-G
1	89.54	70.12
2	90.32	72.74
3	93.29	73.85
4	93.30	73.88

E. Ablation Study

We contrasted our method with two mainstream graph-based approaches, Graph Convolutional Network (GCN) [35] and Graph Attention Network (GAN) [36], with results shown in Table IV. Our hypergraph-based framework consistently outperforms both baselines across both Synapse-G and TotalSeg-G datasets. This performance gap confirms that modeling multi-organ relationships through hyperedges captures richer anatomical context than pairwise graph methods, which are inherently limited to binary interactions.

Question: From the patient's perspective, is the spleen located to the left of the liver?

Answer:

MISS: No (✘)

MISS*: Yes (✔)

Explanation: "Left of" here refers to anatomical left (patient's left), not image-left. The spleen lies in the left upper quadrant, while the liver is primarily in the right upper quadrant.

Fig. 6. Example of one challenge question and corresponding answers. While the baseline model MISS [33] incorrectly answers "No", the enhanced model MISS* on the augmented datasets correctly answers "Yes".

TABLE VI
ABLATION STUDY OF LOSS FUNCTIONS, INCLUDING CROSS-ENTROPY LOSS $\mathcal{L}_{CE}$, CATEGORY-AWARE SUPERVISED CONTRASTIVE LOSS $\mathcal{L}_{Con}$, FOCAL LOSS $\mathcal{L}_{Focal}$, AND OUR DESIGNED GRAPH LOSS. WE UTILIZE F1-SCORE AS METRICS.

Dataset	Synapse-G	TotalSeg-G
$\mathcal{L}_{CE}$	89.71	71.26
$\mathcal{L}_{Con}$	88.72	68.09
$\mathcal{L}_{Focal}$	89.05	71.28
$\mathcal{L}_{CE} + \mathcal{L}_{Con}$	90.34	72.75
$\mathcal{L}_{CE} + \mathcal{L}_{Focal}$	92.57	72.74
Ours	93.30	73.88

In medical knowledge graph generation, the Relation-Aware Hypergraph Module and RA-HA mechanism has a great influence on the performance of multi-organ knowledge graph generation. We consider the influence of RA-HA on the performance of the model, then show the performance with different numbers of RA-HA in Table V. Note that as the number of layers increases, the performance also becomes better on both datasets. The experimental results demonstrate that the performance of the two datasets is best when the number of layers is set to four.

Moreover, we conduct ablation study on different loss functions for medical knowledge graph generation, including cross-entropy loss $\mathcal{L}_{CE}$, category-aware supervised contrastive loss $\mathcal{L}_{Con}$ [37], Focal loss $\mathcal{L}_{Focal}$, and our designed graph loss. The results in Table VI demonstrates that our proposed graph loss achieves the highest F1-scores on both Synapse-G and TotalSeg-G, outperforming all individual and combined losses. This indicates that our graph loss effectively encodes structural and relational priors inherent in anatomical knowledge graphs, enabling the model to generate not only pixel-accurate segmentation but also coherent organ interactions.

V. Limitations and Future Work

While MedGraphGen achieves promising results in generating anatomically plausible knowledge graphs from CT images, several limitations remain. First, the proposed approach is primarily trained and evaluated on CT scans, which can limit its generalizability to other imaging modalities such as MRI and PET. Differences in contrast and resolution might require additional adaptation strategies. Moreover, the multi-organ knowledge graph captures spatial and functional relationships, so it assumes a relatively static anatomical structure. However, inter-organ relationships vary due to pathological conditions, individual anatomical variations, and imaging artifacts. Our future work will focus on reducing reliance on labeled data by leveraging self-supervised and few-shot learning methods to improve model adaptability and enable broader application to datasets with limited annotations. Instead of relying on a predefined structure, future research could explore adaptive graph learning techniques that dynamically adjust connections based on anatomical variations and pathological changes. Furthermore, we intend to develop dynamic graph modeling capabilities to support spatiotemporal reasoning over time-series scans, enabling applications in disease progression analysis and surgical planning.

VI. Conclusion

In this paper, we present MedGraphGen, a unified end-to-end framework for generating interpretable medical knowledge graphs from CT images, advancing the field beyond pixel-level segmentation toward structured anatomical understanding. By integrating a Spatial-Semantic Context Aggregator with a Relation-Aware Hypergraph Module, our model explicitly captures both pairwise spatial relations and higher-order organ interactions through geometry-grounded and semantically weighted message passing. To support training and evaluation, we introduce a new benchmark dataset with dense segmentation masks and expert-annotated relational graphs, enabling rigorous study of structured reasoning in medical imaging. Experiments show that MedGraphGen produces anatomically plausible graphs and achieves state-of-the-art performance in multi-organ relation prediction. Moreover, the generated graphs serve as powerful structured priors that significantly enhance the relational reasoning capabilities of medical vision-language models, improving accuracy on Med-VQA tasks involving spatial and functional queries.

VII. Acknowledgement

This work is supported in part by the MBZUAI-IITD grant and the MBZUAI-WIS Joint Program for Artificial Intelligence Research.

REFERENCES

[1] Wentao Zhu and Yufang Huang, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical Physics*, vol. 46, no. 2, pp. 576–589, 2019.

[2] Alexander Jaus and Constantin Seibold, "Towards Unifying Anatomy Segmentation: Automated Generation of a Full-Body CT Dataset," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024-10, pp. 41–47.

[3] Jakob Wasserthal and Hanns-Christian Breit, "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images," *Radiol Artif Intell*, vol. 5, no. 5, pp. e230024, 2023.

[4] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong, "Self-supervised vision-language pretraining for medial visual question answering," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023, pp. 1–5.

[5] Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*. Munich, Germany, 2015, vol. 5, p. 12.

[6] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro R. A. S. Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, Alan Yuille, and Zongwei Zhou, "AbdomenAtlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking," *Medical Image Analysis*, vol. 97, pp. 103285, 2024-10-01.

[7] Yabo Fu, Yang Lei, Tonghe Wang, Walter J. Curran, Tian Liu, and Xiaofeng Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, 2021.

[8] Li Wang, Lihui Wang, Zixiang Kuai, Lei Tang, Yingfeng Ou, Min Wu, Tianliang Shi, Chen Ye, and Yuemin Zhu, "Progressive Dual Priori Network for Generalized Breast Tumor Segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 9, pp. 5459–5472, 2024.

[9] Toshiyuki Okada, Marius George Linguraru, Masatoshi Hori, Ronald Summers, Noriyuki Tomiyama, and Yoshinobu Sato, *Abdominal Multi-organ CT Segmentation Using Organ Correlation Graph and Prediction-Based Shape and Location Priors*, vol. 16, 2013.

[10] Hejun Huang, Zuguo Chen, Ying Zou, Ming Lu, Chaoyang Chen, Youzhi Song, Hongqiang Zhang, and Feng Yan, "Channel prior convolutional attention for medical image segmentation," *Computers in Biology and Medicine*, vol. 178, pp. 108784, 2024-08.

[11] Adrian V. Dalca, John Guttag, and Mert R. Sabuncu, "Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9290–9299.

[12] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed, "Constrained-CNN losses for weakly supervised segmentation," *Medical Image Analysis*, vol. 54, pp. 88–99, 2019.

[13] Nanyan Shen, Ziyan Wang, Jing Li, Huayu Gao, Wei Lu, Peng Hu, and Lanyun Feng, "Multi-organ segmentation network for abdominal CT images based on spatial attention and deformable convolution," *Expert Systems with Applications*, vol. 211, pp. 118625, 2023.

[14] Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, and Yan Wang, "MagicNet: Semi-Supervised Multi-Organ Segmentation via Magic-Cube Partition and Recovery," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23869–23878.

[15] Guan Wang, Zhimin Li, Qingchao Chen, and Yang Liu, "OED: Towards One-stage End-to-End Dynamic Scene Graph Generation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024-06-16, pp. 27938–27947.

[16] Hengyue Liu and Bir Bhanu, "RepSGG: Novel Representations of Entities and Relationships for Scene Graph Generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2024.

[17] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian, "Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020-06, pp. 211–220.

[18] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian, "Multiscale Spatio-Temporal Graph Neural Networks for 3D Skeleton-Based Motion Prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 7760–7775, 2021.

[19] Michelle M. Li, Kexin Huang, and Marinka Zitnik, "Graph representation learning in biomedicine and healthcare," *Nat. Biomed. Eng*, vol. 6, no. 12, pp. 1353–1369, 2022-12.

[20] Md Mostafijur Rahman and Radu Marculescu, "G-cascade: Efficient cascaded graph convolutional decoding for 2d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7728–7737.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Alexey Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[23] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al., "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[26] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille, "Domain adaptive relational reasoning for 3d multi-organ segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*. Springer, 2020, pp. 656–666.

[27] Ozan Oktay, Jo Schlemper, and Folgoc, "Attention u-net: Learning where to look for the pancreas. arxiv," *arXiv preprint arXiv:1804.03999*, vol. 10, 2018.

[28] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[29] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[30] Xiaohong Huang, Zhifang Deng, Dandan Li, and Xueguang Yuan, "Missformer: An effective medical image segmentation transformer," *arXiv preprint arXiv:2109.07162*, 2021.

[31] Md Mostafijur Rahman and Radu Marculescu, "Medical image segmentation via cascaded attention decoding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6222–6231.

[32] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[33] Jiawei Chen, Dingkang Yang, Yue Jiang, Yuxuan Lei, and Lihua Zhang, "Miss: A generative pre-training and fine-tuning approach for med-vqa," in *International Conference on Artificial Neural Networks*. Springer, 2024, pp. 299–313.

[34] Jiawei Chen, Yue Jiang, Dingkang Yang, Mingcheng Li, Jinjie Wei, Ziyun Qian, and Lihua Zhang, "Can llms' tuning methods work in medical multimodal domain?," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 112–122.

[35] Danfeng Hong, Lianru Gao, Jing Yao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 5966–5978, 2021.

[36] Yanyang Yan, Wenqi Ren, Xiaobin Hu, Kun Li, Haifeng Shen, and Xiaochun Cao, "Srgat: Single image super-resolution with graph attention network," *IEEE Transactions on Image Processing*, vol. 30, pp. 4905–4918, 2021.

[37] Jiankai Li, Yunhong Wang, Xiefan Guo, Ruijie Yang, and Weixin Li, "Leveraging predicate and triplet learning for scene graph generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28369–28379.