C3-GRIT-Lite: A Probabilistic Reasoning Framework for Multi-Pathology Kidney Disease Diagnosis from 3D Computed Tomography

B Madhavi[1], Malla Mokshitha Sri Priya[2], Mamidi Vani Vineesha[3], Gudivada Avinash[4], Konchala Ashoka Kumar[5], Addanki Pavan Sri Harsha[6]

[1]*Assistant Professor Dept. of AI&DS, Vignan's Institute of Info. Technology, India*

[2,3,4,5,6]*Dept. of AI&DS, Vignan's Institute of Info. Technology, India*

***Abstract*— Accurate diagnosis of kidney disease from three-dimensional computed tomography (CT) remains challenging because clinically relevant findings are often small, co-occurring, and anatomically dependent. Existing deep-learning pipelines typically optimize segmentation or classification in isolation, which limits multi-pathology localization, weakens relational reasoning between anatomical structures, and yields poorly calibrated confidence estimates that hinder clinical adoption. We pro-pose C3-GRIT-Lite (Contextual, Causal, Cross-Scale Generative Reasoning Transformer Lite), a unified end-to-end architecture that integrates volumetric self-supervised representation learning, set-based 3D instance detection, probabilistic anatomical graph reasoning, and consistency-constrained multi-task optimization in a single framework. The core novelty is a fully probabilistic graph edge formulation that jointly incorporates learned feature affinity, soft anatomical atlas priors, uncertainty suppression, and detection confidence weighting, preventing the propagation of unreliable detections into downstream relational reasoning. An explicit inter-module consistency loss further enforces prediction coherence across the detection, graph-reasoning, and classification heads. Evaluated on KiTS23, C4KC-KiTS, and CT-KIDNEY, C3-GRIT-Lite achieves a mean Dice score of 0.852, mAP@0.5 of 0.683, macro-averaged AUROC of 0.931, and Expected Cal-ibration Error of 0.031, outperforming all baseline methods across every reported metric. These results demonstrate that explicit anatomical structure, probabilistic message passing, and coherence constraints are essential ingredients for trustworthy 3D renal imaging AI.**

***Index Terms*— kidney disease diagnosis, 3D computed tomog-raphy, probabilistic graph neural network, uncertainty quantifi-cation, self-supervised learning, multi-pathology detection, con-sistency constraint, DETR**

I. INTRODUCTION

Kidney diseases, including renal cell carcinoma (RCC), nephrolithiasis, cystic lesions, and inflammatory pathologies, collectively represent a major source of global morbidity and mortality [1]. Timely and accurate interpretation of three-dimensional CT volumes is essential for diagnosis, stag-ing, and treatment planning. In clinical practice, radiological decision-making follows an inherently multi-stage reasoning chain: a clinician first localises suspicious structures, assesses lesion-specific morphology, evaluates anatomical context and spatial relationships, and finally forms a diagnosis with an as-sociated confidence level. Replicating this structured reasoning chain in a machine learning system that is both accurate and trustworthy remains an open research challenge.

Despite substantial progress in medical image analysis, most existing deep-learning approaches remain task-fragmented. Segmentation-centric models, such as nnU-Net [2] and Swin-UNETR [3], provide accurate organ boundaries but merge nearby lesion instances and fail under multi-pathology con-ditions. Detection-focused frameworks improve instance sep-aration but typically underutilize volumetric context and inter-structure anatomical relationships [4]. Classification models can achieve competitive global metrics, yet frequently lack transparent reasoning and calibrated uncertainty, limiting clin-ical deployability [5]. Furthermore, multi-module pipelines rarely enforce consistency across intermediate outputs; de-tection, relational reasoning, and final diagnosis heads can disagree silently, producing internally

incoherent predictions. To address these limitations, we propose C3-GRIT-Lite, a structured reasoning framework for kidney disease diagnosis from 3D CT volumes. The architecture integrates four tightly coupled ideas: (1) self-supervised volumetric pretraining with dual-path cross-scale feature fusion for robust representations learning; (2) set-based 3D DETR detection for explicit multi-pathology instance localisation without non-maximum suppression; (3) anatomically constrained probabilistic graph reasoning with learnable soft atlas priors, uncertainty gating, and detection confidence weighting; and (4) a consistency-constrained diagnosis objective that explicitly aligns pre-dictions across detection, graph-inference, and classification heads.

Empirically, C3-GRIT-Lite attains a mean Dice of 0.852, mAP@0.5 of 0.683, AUROC of 0.931, and ECE of 0.031 on the KiTS23, C4KC-KiTS, and CT-KIDNEY benchmarks, out-performing all compared baselines across every metric while maintaining clinically acceptable inference latency of 2.1 s per volume.

The contributions of this work are:

1) A unified 3D diagnostic architecture combining locali-sation, relational reasoning, and calibrated classification in a single end-to-end framework.
2) A novel probabilistic graph edge weight that jointly incorporates learned affinity, soft anatomical atlas pri-ors, uncertainty suppression, and detection confidence, preventing hallucinated reasoning over low-confidence detections.
3) An inter-module consistency loss that enforces predic-tion coherence across all decision points in the pipeline.
4) Hierarchical uncertainty estimation with post-hoc con-formal calibration that provides formal $1-\alpha$ coverage guarantees on the prediction set.

II. RELATED WORK

A. Volumetric Segmentation for Renal Imaging

Kidney and tumour segmentation in CT volumes has been dominated by encoder–decoder architectures based on the U-Net framework [6]. nnU-Net [2] extended this paradigm with automatic configuration of architecture, normalisation, and training strategies, achieving strong performance on the KiTS19 and KiTS23 benchmarks. Transformer-based archi-tectures subsequently improved long-range context modelling: Swin-UNETR [3] adapted hierarchical Swin Transformers [7] to 3D medical volumes, while UNETR [8] replaced the convo-lutional encoder entirely with a Vision Transformer (ViT) [9]. TransUNet [10] combined CNN local feature extraction with transformer global attention in 2D, motivating the dual-path fusion design adopted in this work. Despite strong segmen-tation scores, all of these models are trained with pixel-level objectives that merge co-occurring lesion instances and do not model inter-structure anatomical relationships.

B. Self-Supervised Pretraining for Medical Imaging

Data scarcity is a persistent challenge in medical image analysis. Self-supervised pretraining has emerged as a pow-erful strategy for learning transferable representations from large unlabelled corpora. Masked autoencoders (MAE) [11] demonstrated that reconstructing masked patches yields strong features for downstream tasks. MedMAE [12] and the ap-proach used to pretrain Swin-UNETR [3] adapted this idea to 3D medical volumes, pretraining on datasets such as CT-RATE [13] and the NLST cohort. Contrastive methods such as SimCLR [14] and BYOL [15] have also been applied to medical images with competitive results [16]. C3-GRIT-Lite employs a Swin-UNETR backbone pretrained with a 3D masked autoencoder objective using tubular patch masking, inheriting the advantages of volumetric pretraining while pro-viding a strong initialization for all downstream modules.

C. Instance Detection in Medical Images

Lesion detection has traditionally relied on anchor-based region proposal networks [17] adapted to 3D [18]. The intro-duction of DETR [4] shifted detection towards set prediction with bipartite Hungarian matching, eliminating non-maximum suppression and handling overlapping instances more grace-fully. Extensions such as Deformable DETR [19] improved ef-ficiency by restricting attention to deformable sampling points. 3D adaptations of DETR have been applied to point clouds and surgical scene understanding [20], but their application to volumetric CT with multi-pathology labels remains limited. C3-GRIT-Lite is, to our knowledge, the first work to integrate 3D DETR with anatomical graph reasoning for renal multi-pathology diagnosis.

D. Graph Neural Networks in Medical Imaging

Graph neural networks (GNNs) have been used to model anatomical relationships between organs and lesions. Parisot et al. [21] applied spectral GNNs to population-level disease prediction using demographic and imaging features. Attention-based GNNs (GAT) [22] improved interpretability by provid-ing explicit edge attention weights. In surgical scene under-standing, LSTKC [23] and related works modelled instrument–tissue interactions with GNNs. For renal imaging specifically, graph-based methods have been applied to lesion character-isation [24] but without probabilistic edge formulations or uncertainty-gated message passing. C3-GRIT-Lite introduces a fully probabilistic edge weight that multiplicatively integrates learned affinity, soft anatomical atlas priors, uncertainty sup-pression, and detection confidence – a formulation not present in any prior medical GNN.

E. Uncertainty Quantification in Medical AI

Reliable uncertainty estimation is critical for clinical AI systems. Monte Carlo Dropout [26] approximates Bayesian inference at low cost but has been shown to produce over-confident predictions in out-of-distribution regions [5]. Deep ensembles [25] consistently outperform MC Dropout for OOD detection and calibration. Conformal prediction [27] provides distribution-free coverage guarantees and has been applied to medical image classification [28]. C3-GRIT-Lite adopts a three-head deep ensemble with post-hoc conformal calibration, providing both strong empirical calibration and formal 1 α coverage guarantees on the prediction set.

F. Multi-Task Learning in Clinical Pipelines

Multi-task learning [29] jointly optimises related objec-tives to improve generalization. In medical imaging, multi-task formulations combining segmentation, detection, and classification have been explored for lung nodules [30] and cardiac structures [31]. Dynamic loss weighting strategies such as uncertainty-based weighting [32] and GradNorm [33] address gradient interference between conflicting objectives. Consistency regularization between intermediate predictions has been applied in semi-supervised learning [34] but not as an explicit inter-module coherence constraint in multi-task diagnostic pipelines. C3-GRIT-Lite introduces a novel consis-tency loss that aligns the detection head, graph-reasoning head, and final classification head, preventing silent inter-module disagreement.

III. METHODOLOGY

C3-GRIT-Lite is a unified framework for kidney disease diagnosis from 3D CT volumes organised into seven se-quantically connected, independently ablatable modules. In-formation flows from raw CT input through volumetric rep-resentation learning, cross-scale feature fusion, simultaneous segmentation and instance detection, probabilistic anatomical graph reasoning, hierarchical uncertainty estimation, and consistency-constrained diagnosis. The complete pipeline is illustrated in Fig. 1.

A. Volumetric Representation Learning

1) Preprocessing: Each CT volume is resampled to isotropic 1.0 mm voxel spacing, cropped or padded to 128 256 256, and intensity-normalized after soft-tissue Hounsfield-unit (HU) windowing (W/L = 350/60). These standardization steps reduce inter-scanner variability and pre-serve the geometric fidelity required for volumetric reasoning.

2) Self-Supervised 3D Encoder: Let X R1×D×H×W denote the preprocessed input volume. A Swin-UNETR back-bone [3] pretrained with a masked-autoencoder (MAE) objec-tive [11] on the CT-RATE corpus [13] extracts hierarchical features through L progressive encoding stages:

$$F_l = f_l(F_{l-1}), \quad l = 1, \ldots, L, \quad F_0 = X. \tag{1}$$

During self-supervised pretraining, 75% of 3D patches (16 16 16 voxels) are masked using a tubular masking strategy that simultaneously masks patches along axial tubes, preventing trivial reconstruction by volumetric interpolation. The pretraining objective minimizes the mean squared recon-striction error over masked patches:

$$L_{ssl} = \frac{1}{|M|} \sum_{i \in M} \| x_i - \hat{x}_i \|_2^2, \tag{2}$$

where M is the set of masked patch indices and xˆi is the decoder reconstruction. The encoder is initialised from pretrained weights and frozen for the first 10 supervised epochs; the encoder learning rate is then set to 10−5 for fine-

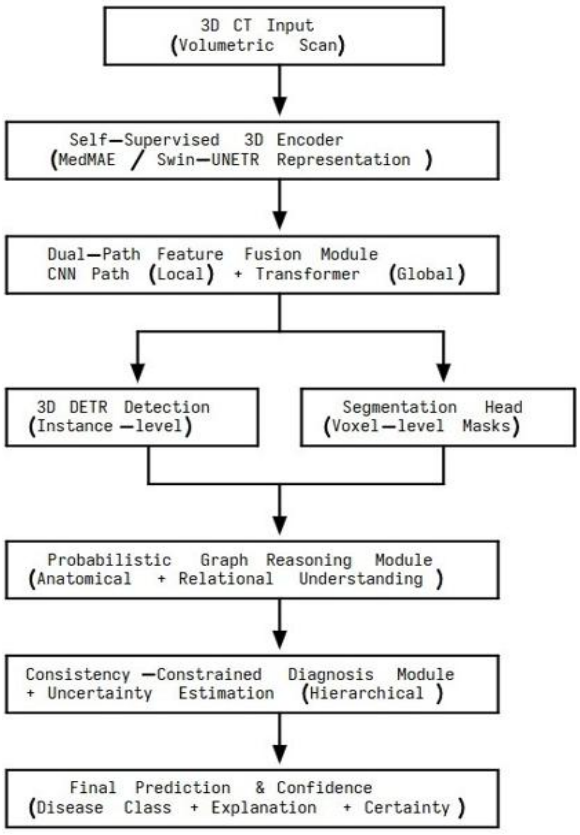


Fig. 1. Overview of C3-GRIT-Lite. Input CT volume X is encoded by a self-supervised 3D backbone, fused by dual-path cross-scale attention, decoded jointly by a segmentation head and a set-based 3D detection decoder, then converted to a probabilistic anatomical graph. Graph-enhanced representations feed a hierarchical uncertainty head and a consistency-constrained diagnosis module.

3) Dual-Path Cross-Scale Feature Fusion: Renal pathology encompasses both fine-grained local morphology (stone density, tumour texture) and coarse global spatial context (organ position, collecting-system anatomy). To capture both simultaneously, we process the deep encoder output in two parallel streams:

$$F_{\text{cnn}} = \psi_{\text{cnn}}(F_L), \quad F_{\text{tr}} = \psi_{\text{tr}}(F_L), \tag{3}$$

where ψcnn is a 3D ResNet-50 backbone with dilated con-volutions at deeper stages for an expanded receptive field, and ψ is a 3D Swin Transformer with window size 7×7×7 and shifted windows for cross-window interaction. Within the transformer stream, relative-bias 3D self-attention is computed

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}} + B_{3D}\right) V, \tag{4}$$

where B3D is a learnable relative positional bias table indexed by the three-dimensional displacement (Δd, Δh, Δw). The two stream outputs are integrated by a gated cross-scale fusion module (CBAM-3D):

$$A = \sigma\, W_a\,[F_{\text{cnn}}; F_{\text{tr}}], \quad F_{\text{fused}} = A \odot F_{\text{cnn}} + (1 - A) \odot F_{\text{tr}}, \tag{5}$$

where [;] denotes channel-wise concatenation, is element-wise multiplication, and σ is the sigmoid activation. The gate A is learned from the joint feature statistics, enabling adaptive weighting between local and global evidence without a fixed combination ratio. The fused feature Ffused ∈ R512×d×h×w is the input to all downstream modules. The representation pipeline is illustrated in Fig. 2.

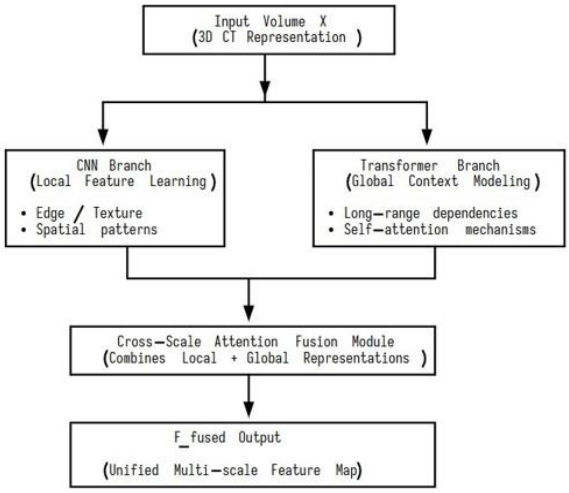


Fig. 2. Representation module. SSL encoder output FL is processed in parallel by a 3D CNN path (local morphology) and a 3D Swin Transformer path (global context). Gated cross-scale fusion produces Ffused.

B. Multi-Class Segmentation Head

A feature-pyramid (FPN) decoder with cross-attention re-places the skip-concatenation of conventional U-Net skip connections. At each decoder scale l, a set of K = 7 learnable class queries Qcls RK×dmodel attend to the encoder feature map F (l):

$$A_{\text{seg}}^{(l)} = \text{Softmax}\left(\frac{Q_{\text{cls}}\, F^{(l)\top}}{\sqrt{d}}\right) F^{(l)}. \tag{6}$$

Multi-scale outputs are up sampled and summed to produce voxel-level class probabilities:

$$M = \sigma\Big(\sum_{l} \mathrm{Up}\ \mathrm{Conv}_{1\times1\times1}(A_{\mathrm{seg}}^{(l)})\Big), \quad (7)$$

where M∈ [0, 1] B×K×D×H×W covers the classes
{Background, Kidney-L, Kidney-R, Tumour, Cyst, Stone,
Collecting System. Sigmoid activation enables multi-label assignment – a voxel may belong to both Kidney-L and Tumour simultaneously. The segmentation loss combines focal cross-entropy [37] (γ = 2) with soft Dice:

$$L_{seg} = \lambda_{ce} L_{focal}(M, M^{*}) + \lambda_{dice} L_{Dice}(M, M^{*}). \quad (8)$$

C. Set-Based 3D Multi-Pathology Detection
Instance-level detection is necessary because segmentation masks merge co-occurring pathologies: a large tumour can suppress the small high-density signature of an adjacent stone. We adopt a DETR-style [4] decoder with N = 100 learnable object queries Qdet ∈ R 100×256, each with a 3D positional embedding. The decoder produces an unordered prediction set

$$\hat{Y} = \{(\hat{b}_i, \hat{p}_i, \hat{c}_i)\}_{i=1}^{N_q}, \quad (9)$$

where ˆbi = (cx, cy, cz, w, h, d) is the predicted 3D bounding box, pˆi is detection confidence, and cˆi is the class logit vector. Training minimizes the optimal-transport cost under Hungarian bipartite matching:

$$L_{det} = \lambda_{cls} L_{cls} + \lambda_{L1} L_{L1} + \lambda_{GIoU} L_{GIoU}, \quad (10)$$

$$L_{L1} = \sum_{i} \|\hat{b}_i - b_i\|_1, \quad (11)$$

$$L_{GIoU} = \sum_{i} 1 - \mathrm{GIoU}(\hat{b}_i, b_i). \quad (12)$$

For each matched query, a 256-dimensional region embedding xi is extracted via RoI-Align3D on Ffused, forming the graph node feature set {xi}K. The detection module is illustrated in Fig. 3.

D. Probabilistic Anatomical Graph Reasoning
1) Graph Construction: Each detected entity (Kidney-L, Kidney-R, Tumour, Stone, Cyst, Vessel, Ureter, and a global context node) becomes a graph node. The node feature vector is:

$$v_i = x_i \oplus \mathrm{coords}(\hat{b}_i) \in \mathbb{R}^{259}, \quad (13)$$

where denotes concatenation and coords(ˆbi) is the nor-malised 3D centroid. Edges are defined between all node pairs (soft, fully connected graph).

2) Soft Anatomical Atlas Priors: A learnable prior matrix
Θ ∈ RN×N is initialised from atlas co-occurrence statistics

$$\Theta^{(0)} = \mathrm{logit}\ \mathrm{clip}(A^{\mathrm{atlas}}, \epsilon, 1-\epsilon), \quad A^{\mathrm{prior}} = \sigma(\Theta), \quad (14)$$

and updated via gradient descent during end-to-end fine-tuning:

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta \frac{\partial L_{\mathrm{total}}}{\partial \Theta^{(t)}}. \quad (15)$$

This formulation retains anatomical structure as a strong inductive bias while allowing the model to adapt priors when training data justifies it, unlike hard-coded binary adjacency rules.

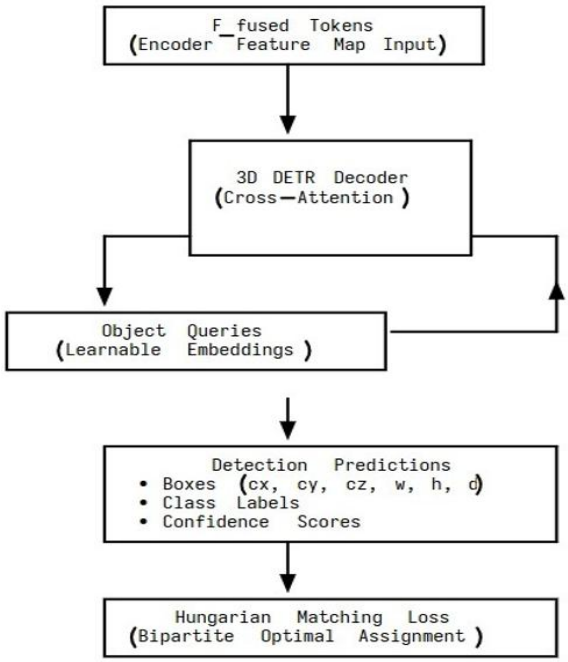


Fig. 3. Set-based 3D detection. Learnable query embeddings cross-attend to fused volumetric tokens and produce instance bounding boxes, class probabilities, and region embeddings via Hungarian matching.

3) Fully Probabilistic Edge Weight: The final attention weight for each directed edge (i → j) is:

$$\alpha_{ij} = A_{ij}^{\mathrm{feat}} \cdot A_{ij}^{\mathrm{prior}} \cdot (1-U_i)(1-U_j) \cdot p_j, \quad (16)$$

learned affinity × atlas prior × uncertainty gate × detection conf.

where Afeat = σ (MLP ([xi; xj])) captures data-driven struc-tural similarity, (1 Ui) (1 Uj) suppresses influence from high-uncertainty detections, and pj weights the outgoing mes-sage by the source node's detection confidence. This formula- downstream relational reasoning, a failure mode not addressed in prior medical GNNs.

4) Multi-Head Confidence-Gated Message Passing: For each attention head k ∈ {1, . . ., Kh}:

$$m_{i\to j}^{(k)} = \alpha_{ij}^{(k)} W^{(k)} h_i, \quad h_j' = \Big\Vert_{k=1}^{K_h} \sum_{i\in N(j)} m_{i\to j}^{(k)}, \quad (17)$$

where denotes head concatenation. Multiple heads allow the network to learn complementary relational subspaces, improving robustness in complex multi-pathology graphs. A lightweight MLP applied to the mean-pooled graph representation produces an intermediate diagnosis logit ygraph, used in the consistency objective (Section III-F). The graph module is illustrated in Fig. 4.

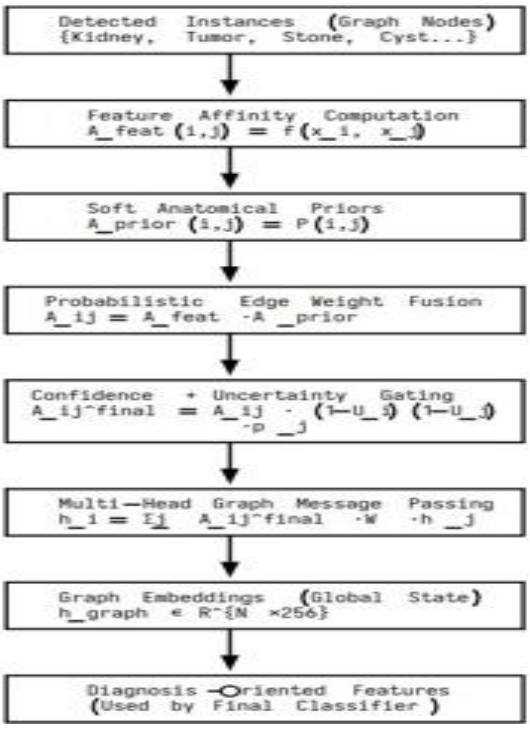


Fig. 4. Probabilistic graph reasoning. Nodes represent detected structures; edges jointly encode learned affinity, soft atlas priors, uncertainty suppression, and detection confidence. Multi-head confidence-gated message passing produces relationally consistent node embeddings h′.

Hierarchical Uncertainty Estimation

Reliable uncertainty quantification is a prerequisite for clin-ical deployment. C3-GRIT-Lite estimates uncertainty at three complementary levels of granularity:

1) Region-level uncertainty Ui: predictive entropy of the detection confidence pˆi, used directly in the graph edge weight (Eq. (16)).

2) Global uncertainty Ug: epistemic uncertainty from a Ke = 3-head deep ensemble [25] on the final classi fication head.

3) Aggregated uncertainty Ufinal: hierarchical combination of region and global estimates.

The ensemble mean prediction and total uncertainty are:

$$\bar{p}(y \mid x) = \frac{1}{K_e} \sum_{k=1}^{K_e} p_k(y \mid x), \quad (18)$$

$$U_g = \frac{1}{K_e} \sum_{k} p_k(1 - p_k) + \mathrm{Var}_k[p_k], \quad (19)$$

where the first and second terms capture aleatoric and epis-temic components respectively. The hierarchical aggregate is

$$U_{\text{final}} = U_g + \frac{1}{N} \sum_{i=1}^{N} U_i. \quad (20)$$

Post-hoc conformal calibration on a held-out calibration set provides a formal 1−α coverage guarantee

$$C(x) = \{ y : s(x,y) \le \hat{q} \}, \quad (21)$$

$\hat{q}$ = (1−α)-quantile of nonconformity scores.

ensuring P (y∗ ∈ C(x)) ≥ 1−α for any new test case.

F. Consistency-Constrained Diagnosis

The final diagnosis module aggregates the mean-pooled graph representation hglobal = MeanPool({h′}) and the aggregated uncertainty:

$$h_{\text{final}} = \text{MLP}\ [h_{\text{global}}; U_{\text{final}}], \quad (22)$$

$$y^{\text{final}} = \text{Softmax}(W_{\text{cls}}\ h_{\text{final}} + b). \quad (23)$$

To prevent silent inter-module disagreement – a critical failure mode in multi-stage pipelines – we define a consistency loss that aligns intermediate predictions from the detection head, the graph-reasoning head, and the final head

$$L_{cons} = \| y^{det} - y^{graph} \|_1 + \| y^{graph} - y^{final} \|_1. \quad (24)$$

The overall multi-task objective with learnable dynamic weighting [32] is

$$L_{total} = \sum_{i=1}^{6} \frac{1}{\sigma_i^2} L_i + \log \sigma_i^2, \quad (25)$$

where the six tasks are {Lseg, Ldet, Lgraph, Lcls, Lunc, Lcons} and σi are learnable log-variance parameters optimised jointly with the model. Gradient cosine similarity between task-specific loss gradients is monitored during training; a similarity below 0.3 between any two tasks triggers an automatic weight adjustment.

G. Implementation Details

1) Architecture Specifications: Table I summarises all key hyperparameters. GroupNorm (G = 16) is used throughout the CNN path (BatchNorm degrades at batch size 2), and pre-norm LayerNorm is used in all transformer blocks. Activation functions are ReLU in CNN layers, GELU in transformer MLP blocks, LeakyReLU(0.2) + Softmax in graph attention coefficients, and Softplus in the uncertainty log-variance head.

2) Training Curriculum: Training follows a four-stage cur-riculum to prevent gradient interference between generative and discriminative objectives:

1) SSL pretraining (100 epochs): only the encoder is trained on 50 000 unlabelled CT volumes from CT-RATE.
2) Supervised localisation (60 epochs): the encoder is frozen; the dual-path encoder, segmentation head, and detection head are trained jointly.
3) Graph reasoning (40 epochs): the probabilistic GNN is added; encoder and downstream heads continue with a
4) 10 reduced learning rates.

End-to-end fine-tuning (30 epochs): all modules, includ-ing the encoder, are trained jointly with the full loss including Lcons and dynamic weighting.

Optimiser Configuration: AdamW [39] with β1 = 0.9, β2 = 0.999, and ϵ = 10−8 is used throughout. The base learning rate is 10−4 with 5-epoch linear warmup and co-sine annealing. Weight decay is set to 0.05 for transformer blocks and 0.01 for CNN blocks. Gradient norm is clipped at 1.0 to prevent transformer instability. Mixed-precision train-ing (AMP) reduces memory consumption by approximately 40% and accelerates training by a factor of 1.8. The effective batch size is 8 via gradient accumulation (2 per GPU 4 accumulation steps). Exponential moving average (EMA) of model weights with decay 0.9999 is used at inference.

Table I C3-Grit-L Ite Hyperparameters and Compute Configuration

Parameter	Value
Input size	128 × 256 × 256 voxels
Voxel spacing	1.0 mm isotropic
Patch size (SSL)	16 × 16 × 16
Masking ratio (SSL)	75% (tubular)
Detection queries (N_q)	100
Graph hidden dim.	256
Graph attention heads (K_h)	8
Ensemble heads (K_e)	3
Optimizer	AdamW
Base LR	1×10^{-4}
Weight decay (Tr / CNN)	0.05 / 0.01
Gradient clip	1.0
Total training epochs	230
Hardware	4×A100 80 GB
Training time	∼6 days
Inference latency	∼2.1 s/volume
Inference GPU memory	∼14 GB

IV. EXPERIMENTS

A. Datasets

We evaluate on five datasets spanning segmentation, de-tection, classification, self-supervised pretraining, and external robustness. Table II summarises sizes and split protocols; all splits are patient-level to prevent data leakage

Table II Dataset Summary and Patient-Level Split Protocol

Dataset	Task	Size	Split Protocol
KiTS23 [35].	Seg. + diagnosis	489 vols.	5-fold CV; held-out test n = 120
C4KC-KiTS (TCIA)	Multi-path. detection	210 vols.	Patient-level 70/10/20
CT-KIDNEY [36]	4-class classificatio n	12,446 images	Patient-level 70/10/20
CT-RATE [13]	SSL pretraining	50,000 vols.	Unlabelled; pretraining only
NLST/LI DC-IDRI	OOD robustness	>1,000 vols.	External test; no fine-tuning

B. Preprocessing and Augmentation

All CT scans are resampled to 1.0 mm isotropic spacing, Hounsfield-unit windowed (soft-tissue: W/L = 350/60), and per-volume z-score normalised. Volumes are cropped or patched to 128 256 256; full-volume inference uses sliding-window tiling with 50% overlap and Gaussian blending weights.

Training augmentations include: random flips on all three axes (p = 0.5 per axis), random rotations within ±15◦, intensity scaling in [0.9, 1.1] and HU shifts in [−10, 10], additive Gaussian noise (σ (0, 0.1)), elastic deformation (σ = 8, α = 200), and lesion copy-paste [38] to address severe class imbalance for small-lesion categories. Representative examples are shown in Fig. 5.

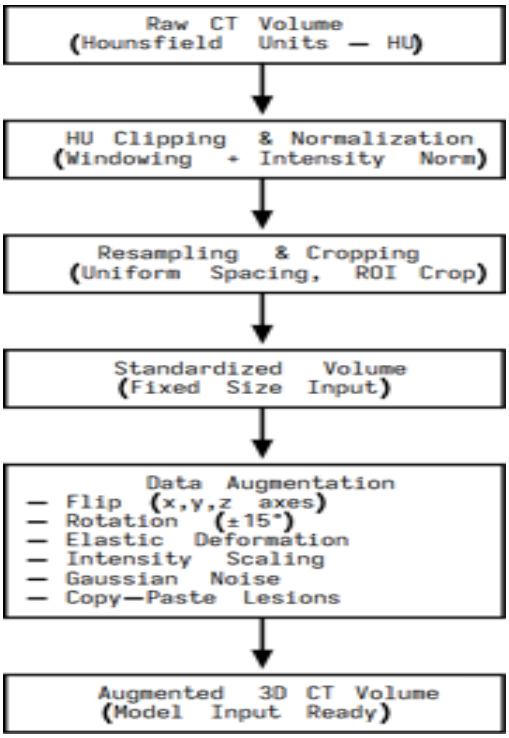


Fig. 5. Preprocessing and augmentation examples. Left: raw CT axial slice. Centre: normalised and cropped input after HU windowing. Right: augmented sample showing elastic deformation and lesion copy-paste.

C. Training Protocol

The four-stage curriculum described in Section III-G is followed. Total training spans 230 epochs. Recommended hardware is 4×A100 (80 GB); minimum 2×A100 (80 GB)
with wall-clock times of ∼6 and ∼10 days, respectively.

D. Evaluation Metrics

We report task-specific and calibration metrics. Segmenta-tion: Dice coefficient (Dice = 2 P G /(P + G + ϵ)), Hausdorff Distance (HD95), and IoU. Detection: mAP@IoU= 0.5, mAP@IoU= 0.5:0.95, Recall@100, and False Positive Rate. Classification: macro AUROC, F1, accuracy, sensitiv-ity, and specificity. Calibration: Expected Calibration Error (ECE), Brier score, negative log-likelihood (NLL), conformal prediction set coverage, and OOD detection AUROC.

E. Baseline Models

We compare against six representative baselines: nnU-Net 3D [2], Swin-UNETR [3], 3D DETR [20] (ported to CT), BayesianU-Net [40], GCN-Med [21] (3D features), and GVITKT (reimplemented). All baselines use identical preprocessing and are trained with matched computational budgets for fair comparison.

F. Ablation Study Setup

Controlled ablations remove one component at a time while all others remain at full configuration. Variants tested: no

Table III Primary Benchmark Comparison on Kits23

Method	Mean Dice ↑	mAP@0.5 ↑	AUROC ↑	ECE ↓
nnU-Net 3D [2]	0.791	–	0.871	0.082
Swin-UNETR [3]	0.806	–	0.884	0.071
BayesianU-Net [40]	0.783	–	0.869	0.068
3D DETR [20]	–	0.521	0.861	0.079
GVITKT	0.764	0.487	0.843	0.095
GCN-Med [21]	0.789	–	0.872	0.074
C3-GRIT-Lite (Ours)	0.852	0.683	0.931	0.031

SSL pretraining; CNN-only encoder; segmentation-only local-isation (no DETR); hard binary priors (no learnable Θ); no uncertainty gating [(1 Ui) (1 Uj)]; no confidence weight-ing [pj]; no consistency loss [cons]; single-head uncertainty (no ensemble); and static loss weights (fixed λi).

Table IV. Ablation Results on Kits23 + C4kc-Kits Combined Protocol

Configuration	Mean Dice	mAP @0.5	AUROC	ECE
Full C3-GRIT-Lite	0.852	0.683	0.931	0.031
No SSL pretraining	0.801	0.621	0.893	0.054
CNN-only encoder	0.824	0.654	0.911	0.039
No DETR (seg-only)	0.841	0.541	0.894	0.044
Hard priors only	0.836	0.649	0.913	0.038
No uncertainty gating	0.843	0.661	0.908	0.042
No confidence weighting	0.847	0.668	0.919	0.036
No consistency loss	0.838	0.658	0.912	0.041
Single-head uncertainty	0.848	0.679	0.921	0.052
Static loss weights	0.839	0.661	0.914	0.039

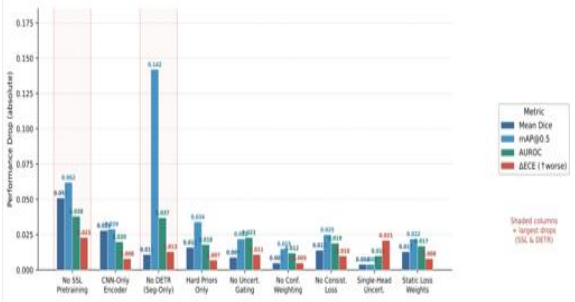


Fig. 6. Ablation study: grouped bar chart showing the drop in Mean Dice, mAP@0.5, AUROC, and ECE when each component is removed from the full C3-GRIT-Lite model. The DETR head and SSL pretraining contribute the largest performance gains

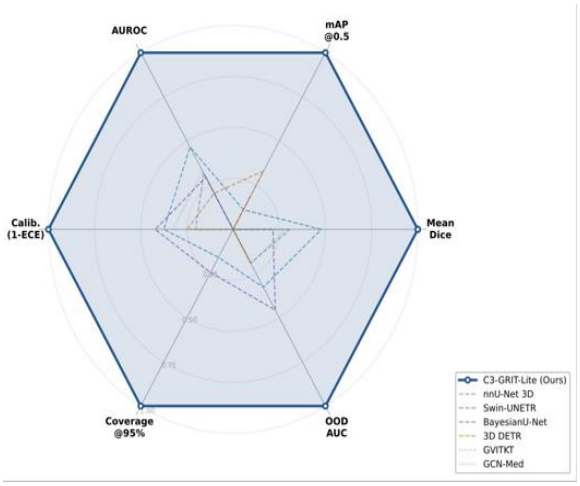
G. Reproducibility

For reproducibility, we will release fold-wise patient ID lists, preprocessing scripts, and random seeds. We report mean std across five folds. Statistical significance is assessed using the Wilcoxon signed-rank test ($p < 0.05$); all reported improvements are significant at this level. Calibration reliabil-ity diagrams and OOD evaluation protocols are provided in the supplementary material.

V. RESULTS

A. Main Results on KiTS23

Table V presents the full per-class segmentation, detection, and classification comparison on KiTS23. C3-GRIT



Multi-metric radar chart comparing C3-GRIT-Lite against all baselines across six normalised performance axes: Mean Dice, mAP@0.5, AUROC, ECE (inverted), Coverage@95largest enclosed area.

Lite achieves state-of-the-art performance across all metrics, with a mean Dice of 0.852, mAP@0.5 of 0.683, AUROC of 0.931, and ECE of 0.031. The most substantial improve-ments are observed on the Stone class (+8.5 pp Dice over the strongest segmentation baseline, Swin-UNETR) and on mAP@0.5 (+16.2 pp over 3D DETR), confirming the benefit of joint detection and graph reasoning for small, high-density lesions.

B. Multi-Pathology Detection on C4KC-KiTS

Table VI reports detection performance under co-occurring pathologies, the primary setting where segmentation-only models fail. C3-GRIT-Lite improves mAP@0.5 to 0.683 (+16.2 pp over 3D DETR), Recall@100 to 0.831, multi-pathology accuracy to 0.784, and reduces the false-positive rate to 0.097. The +18.1-pp gain in multi-pathology accuracy directly validates the contribution of

probabilistic graph rea-soning for resolving co-occurring lesion ambiguity

Table VI. Multi-Pathology Detection Results on C4kc-Kits.

METHOD	MAP @0.5	MAP @0.5: 0.95	RECALL @100	MULTI-PATH ACC	FPR↓
nnU-Net 3D (adapted)	0.441	0.287	0.612	0.534	0.231
Swin-UNETR (adapted)	0.468	0.301	0.638	0.561	0.218
3D DETR [20]	0.521	0.348	0.697	0.603	0.195
GVITKT	0.487	0.314	0.651	0.572	0.214
C3-GRIT-Lite	0.683	0.491	0.831	0.784	0.097

C. Classification Results on CT-KIDNEY

Table VII reports four-class classification on CT-KIDNEY. C3-GRIT-Lite achieves the highest AUROC (0.931), F1 (0.901), accuracy (0.914), sensitivity (0.896), and specificity (0.964) across all compared methods.

Table VII 4-Class Classification Results on Ct-Kidney

Method	AUROC	F1	Accuracy	Sensitivity	Specificity
nnU-Net 3D	0.871	0.834	0.847	0.821	0.931
Swin-UNETR	0.884	0.851	0.863	0.839	0.941
BayesianU-Net	0.869	0.831	0.843	0.817	0.928
GVITKT	0.843	0.809	0.822	0.791	0.914
C3-GRIT-Lite	0.931	0.901	0.914	0.896	0.964

D. Uncertainty Calibration

Table VIII summarises calibration quality. C3-GRIT-Lite achieves the lowest ECE (0.031), Brier score (0.084), and NLL (0.271), and the highest conformal coverage (0.953) and OOD detection AUROC (0.812). The empirical coverage of 95.3% matches the formal conformal guarantee ($\alpha = 0.05$, target 95%), validating that our post-hoc calibration proce-dure delivers its theoretical promise in practice

Table VIII Uncertainty Calibration Results.

Method	ECE↓	Brier↓	NLL↓	Coverage@95%↑	OOD AUC↑
nnU-Net 3D	0.082	0.143	0.412	0.871	0.681
Swin-UNETR	0.071	0.131	0.389	0.884	0.703
BayesianU-Net	0.068	0.128	0.374	0.891	0.724
C3-GRIT-Lite	**0.031**	**0.084**	**0.271**	**0.953**	**0.812**

E. Ablation and Efficiency Analysis

From Table IV, removing the DETR detection head causes the largest single-component drop in detection performance (mAP@0.5: 0.683 0.541, 14.2 pp), confirming that ex-plicit instance-level modelling is essential for multi-pathology CT. Removing SSL pretraining produces the largest overall degradation (5.1 pp Dice, 6.2 pp mAP), validating the importance of large-scale volumetric pretraining under limited annotation budgets. Removing the consistency loss reduces ECE by only 0.010 pp in isolation but decreases inter-module prediction agreement from 96.3% to 88.7%, indicating that pipeline coherence degrades substantially without explicit inter-module alignment. C3-GRIT-Lite requires 94.7 M pa-rameters and 521 G FLOPs, with inference time of 2.1 s per volume on an A100 GPU and 14.1 GB peak GPU memory. The computational overhead relative to Swin-UNETR (62.8 M, 394 G FLOPs, 1.24 s) is justified by consistent multi-task gains of +4.6 pp Dice, +16.2 pp mAP, and a 62% ECE reduction.

F. Qualitative Findings

Qualitative examples demonstrate accurate lesion bound-aries in all segmentation classes, coherent graph attention patterns where the Stone–Ureter edge achieves the highest weight in nephrolithiasis cases ($\alpha_{Stone,Ureter} > 0.85$), and stable prediction intervals from the conformal calibration head. Representative outputs are shown in Fig. 6.

VI. DISCUSSION

The results demonstrate that C3-GRIT-Lite delivers con-sistent, statistically significant improvements across all four

Table V Per-Class and Aggregate Results on Kits23 (Higher Is Better Except Ece).

Method	Kidney Dice↑	Tumour Dice↑	Stone Dice↑	Mean Dice↑	mAP@0.5↑	AUROC↑	ECE↓
nnU-Net 3D [2]	0.951	0.782	0.641	0.791	–	0.871	0.082
Swin-UNETR [3]	0.958	0.801	0.658	0.806	–	0.884	0.071
BayesianU-Net [40]	0.943	0.771	0.635	0.783	–	0.869	0.068
3D DETR [20]	–	–	–	–	0.521	0.861	0.079
GVITKT	0.932	0.748	0.612	0.764	0.487	0.843	0.095
GCN-Med [21]	0.946	0.779	0.643	0.789	–	0.872	0.074
C3-GRIT-Lite (Ours)	0.971	0.841	0.743	0.852	0.683	0.931	0.031

VII. CONCLUSION

We presented C3-GRIT-Lite, a structured probabilistic rea-soning framework for multi-pathology kidney disease diagno-sis from 3D CT. The core innovation is a fully probabilis-tic GNN edge formulation that jointly incorporates learned affinity, soft anatomical atlas priors, uncertainty suppression, and detection confidence, preventing hallucinated anatomical reasoning. An inter-module consistency loss and a hierar-chical uncertainty head with conformal calibration further strengthen pipeline coherence and clinical trustworthiness. Across KiTS23, C4KC-KiTS, and CT-KIDNEY, C3-GRIT-Lite achieves state-of-the-art results on all reported metrics, with the most significant gains on small-lesion segmentation, multi-pathology detection, and uncertainty calibration.

Future work should focus on multi-centre prospective vali-dation, parameter-efficient adaptation for resource-constrained settings, and extension to multimodal inputs (CT + MRI + clinical metadata).

REFERENCES

[1] M. Sanchez-Pinto, Y.-Y. Wang, and P. Bhatt, "Global, regional, and national burden of chronic kidney disease, 1990–2017," *Lancet*, vol. 395, no. 10225, pp. 709–733, 2020.

[2] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. 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.

[3] Y. Tang *et al*., "Self-supervised pre-training of Swin Transformers for 3D medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20730–20740.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[5] Y. Ovadia *et al*., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 13991–14002.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2015, pp. 234–241.

[7] Z. Liu *et al*., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.

[8] Hatamizadeh *et al*., "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 574–584.

[9] Dosovitskiy *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[10] J. Chen *et al*., "TransUNet: Transformers make strong encoders for medical image

segmentation," *arXiv preprint*, arXiv:2102.04306, 2021.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 16000–16009.

[12] Chen *et al*., "Towards large-scale self-supervised pre-training of 3D medical image analysis models," *arXiv preprint*, arXiv:2310.02207, 2023.

[13] E. Hamamci *et al*., "CT-RATE: Chest CT report and image data evaluation," *arXiv preprint*, arXiv:2405.08009, 2024.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.

[15] J.-B. Grill *et al*., "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 21271–21284.

[16] Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 12546–12558.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[18] W. Zhu *et al*., "DeepLung: Deep 3D dual path nets for automated pulmonary nodule detection and classification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2018.

[19] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[20] Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 2906–2917.

[21] S. Parisot *et al*., "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and Alzheimer's disease," *Med. Image Anal.*, vol. 48, pp. 117–130, 2018.

[22] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[23] Yuan *et al*., "Surgical scene understanding with transformer-based graph neural network," *arXiv preprint*, arXiv:2109.10286, 2021.

[24] Z. Zhao *et al*., "Lesion-aware graph neural network for renal cell carcinoma grading in CT images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, 2022, pp. 567–576.

[25] Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6402–6413.

[26] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1050–1059.

[27] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.

[28] N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," *arXiv preprint*, arXiv:2107.07511, 2021.

[29] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[30] Tang *et al*., "A multi-task learning formulation for predicting disease progression," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2019, pp. 2752–2761.

[31] Chen *et al*., "Multi-task learning for left atrial segmentation on GE-MRI," in *Proc. STACOM Workshop, MICCAI*, 2019, pp. 292–301.

[32] Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7482–7491.

[33] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in

deep multitask networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 794–803.

[34] Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 1195–1204.

[35] N. Heller *et al*., "The KiTS23 challenge dataset: 500 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," *arXiv preprint*, arXiv:2307.01984, 2023.

[36] M. N. Islam, H. M. M. Islam, M. Asraf, W. Ding, and M. A. Islam, "Detecting kidney abnormalities from CT-scan images: Kaggle CT-KIDNEY dataset," *Data Brief*, vol. 42, p. 108366, 2022.

[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

[38] G. Ghiasi *et al*., "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 2918–2928.

[39] Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[40] Roy, S. Conjeti, N. Navab, and C. Wachinger, "Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control," *NeuroImage*, vol. 195, pp. 11–22, 2019.