Beyond Isolated Scans: Cross-Phase Alignment of Structure and Topology for 3D Medical Pretraining

Wenzhuo Xu[1,2*], Yan-Jie Zhou[1,2*], Yujian Hu[3,4], Hongkun Zhang[4], and Minfeng Xu[1,2†]

[1] DAMO Academy, Alibaba Group, Hangzhou, China
[2] Hupan Laboratory, Hangzhou, China
[3] School of Medicine, Zhejiang University, Hangzhou, China
[4] Department of Vascular Surgery, The First Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, China
{wenzhuo.xwz, zhouyanjie.zyj, eric.xmf}@alibaba-inc.com

Abstract. Self-supervised learning (SSL) has significantly advanced 3D medical image analysis. However, existing pretraining methods typically treat multi-phase scans as isolated samples, neglecting the rich anatomical correspondence inherent in paired acquisitions. Clinically, contrast-enhanced CT (CECT) offers a structure-amplified view of the exact anatomy captured in a paired non-contrast CT (NCCT), yet the drastic intensity shifts caused by contrast agents make direct voxel-wise alignment challenging. To harness this naturally occurring supervisory signal while decoupling structural topology from phase-specific appearance, we propose **CAST** (**C**ross-phase **A**lignment of **S**tructure and **T**opology), a novel paired pretraining framework. CAST employs a 3D CNN architecture to explicitly align NCCT representations with CECT targets. Moving beyond conventional reconstruction, we introduce two feature-level constraints: (1) **a Spectral Consistency module** that utilizes 3D wavelet decomposition to align frequency-aware boundaries while suppressing contrast-induced noise; (2) **a Geometry-Aware Topological Consistency module** that preserves local relational graphs among salient anatomical keypoints via dynamic top-hat sampling. To support this, we construct a large-scale dataset comprising 13,850 paired volumetric CT scans. Extensive evaluations on public benchmarks and real-world clinical cohorts demonstrate that CAST achieves state-of-the-art performance. Notably, our method delivers substantial gains in boundary-sensitive segmentation tasks and low-data regimes, proving its efficacy in learning robust, phase-invariant anatomical representations for both NCCT and CECT downstream applications. Code is available at `https://github.com/alibaba-damo-academy/CAST`.

Keywords: Self-supervised pretraining · Cross-Phase Alignment · Spectral & topological consistency

*These authors contributed equally. †Minfeng Xu is corresponding author.

1 Introduction

Self-supervised learning (SSL) has emerged as a powerful paradigm in computer vision to mitigate the heavy reliance on costly manual annotations [4, 5, 11, 29]. By learning general-purpose representations from large-scale unlabeled data, various SSL approaches have subsequently demonstrated profound success in 3D medical image analysis, capturing rich anatomical contexts without requiring tedious voxel-wise labels [14,31,33,35,40]. Collectively, these studies suggest that anatomical priors can be implicitly learned from large-scale unlabeled volumes, laying the foundation for data-efficient medical representation learning.

To further push the boundaries of these volumetric representations, it is crucial to exploit the rich, intrinsic supervisory signals naturally embedded in clinical data. Specifically, routine clinical workflows frequently acquire multi-phase or multi-contrast 3D volumes (e.g., unenhanced and contrast-enhanced scans) for the same patient within a short interval [2, 20]. Although contrast agents and protocol variations can drastically shift voxel intensity distributions, the underlying anatomical topology remains largely consistent across phases, with only minor misalignments arising from motion and acquisition variability [8, 17]. Consequently, these multi-phase acquisitions provide structure-amplified observations of the exact anatomy, creating a massive, untapped reservoir of naturally paired data for self-supervised pretraining.

Despite the availability of this natural correspondence, most existing SSL frameworks remain phase-agnostic, typically treating multi-phase scans as isolated, independent samples [13, 32]. Relying heavily on standard reconstruction or global matching objectives, these methods inherently encourage models to memorize phase-dependent intensity patterns rather than abstracting robust anatomical features [14]. A key challenge is that contrast agents dramatically alter the absolute intensity of structural boundaries. Directly aligning spatial features across phases forces the network to hallucinate these artificial brightness shifts. Furthermore, simple global alignment ignores the fact that contrast agents selectively highlight specific physiological structures (e.g., vascular beds), which naturally serve as highly salient anatomical landmarks. As illustrated in Fig. 1 (left), instead of treating these multi-phase scans as isolated inputs, we argue that paired cross-phase 3D data offers a unique opportunity for structure-centered representation learning. By treating the enhanced scan, which offers clear local topological relationships, as a privileged target, we can explicitly decouple structural topology from phase-specific texture.

In this work, we bridge multi-phase clinical acquisition and self-supervised representation learning by proposing **CAST** (**C**ross-phase **A**lignment of **S**tructure and **T**opology). Unlike conventional phase-agnostic approaches, CAST leverages contrast-enhanced CT (CECT) as a *structure-amplified view* to guide representation learning for non-contrast CT (NCCT). While contrast agents significantly alter voxel intensities, the underlying anatomical topology and structural boundaries remain largely preserved. To exploit this property, we introduce two complementary constraints beyond standard reconstruction: (1) a *Spectral Consistency* module that aligns frequency-aware structural information via 3D wavelet de-

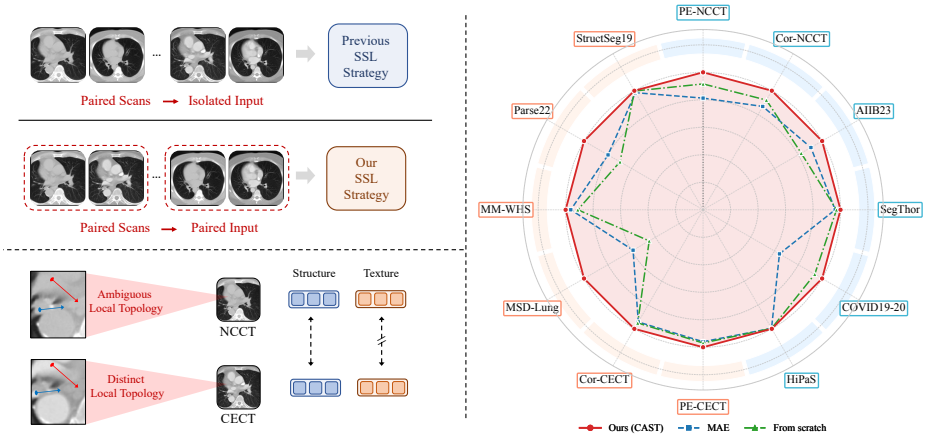

Fig. 1: Motivation and broad applicability of the proposed CAST framework. (Left) While conventional self-supervised learning (SSL) strategies treat multi-phase scans as isolated inputs, CAST explicitly leverages their naturally paired correspondence. By decoupling phase-specific appearance (texture) from the underlying anatomy (structure), our method utilizes the clear local topological relationships inherent in contrast-enhanced scans to guide the representation learning of unenhanced inputs. (Right) Comprehensive evaluation across 12 diverse downstream datasets. CAST (red solid line) consistently outperforms the strong baseline (blue dashed line) on both Non-Contrast CT (NCCT) and Contrast-Enhanced CT (CECT) tasks, demonstrating superior phase-invariant robustness and generalization.

composition, and (2) a *Geometry-Aware Topological Consistency* module that preserves local relational structures through dynamic graph modeling. Together, these mechanisms encourage phase-invariant anatomical representation learning rather than appearance memorization.

To facilitate this research and address the scarcity of paired data, we curate and release a **large-scale dataset comprising 13,850 paired NCCT-CECT volumetric scans**, representing one of the most extensive collections of its kind for cross-phase pretraining. Extensive experiments across diverse chest CT benchmarks demonstrate consistent performance gains, particularly in boundary-sensitive segmentation and low-label regimes, highlighting the robustness and transferability of the learned structural representations.

Our contributions are summarized as follows:

- We propose CAST, the first self-supervised framework to explicitly leverage CECT as a structure-amplified privileged view for guiding NCCT representation learning. This transforms naturally occurring multi-phase clinical data into robust supervisory signals for learning phase-invariant anatomical features.
- We introduce two complementary feature-level constraints to decouple anatomy from contrast shifts: **Spectral Consistency** via 3D wavelet decomposi-

tion for boundary alignment, and **Geometry-Aware Topological Consistency** via dynamic graph modeling for structural preservation. This effectively mitigates intensity discrepancies while retaining topological integrity.
– We curate and release a massive dataset comprising 13,850 paired NCCT-CECT volumetric scans, addressing the scarcity of aligned multi-phase data. This resource represents one of the largest collections of its kind, establishing a new benchmark for cross-phase pretraining research.
– Extensive experiments on public benchmarks and real-world clinical cohorts demonstrate that CAST achieves leading performance. Notably, CAST delivers substantial gains in boundary-sensitive segmentation tasks and low-data regimes, proving its efficacy in learning robust, phase-invariant anatomical representations for both NCCT and CECT downstream applications.

2 Related Works

2.1 Self-Supervised Learning in 3D Medical Imaging

Self-supervised learning (SSL) has emerged as a pivotal strategy for leveraging large-scale unlabeled 3D medical volumes. Early frameworks like Models Genesis [40] established the efficacy of using intrinsic anatomical patterns as supervision signals. Recent advancements have predominantly focused on Masked Image Modeling (MIM). Methods such as HybridMIM [37] and MiM [41] employ hierarchical and multi-scale masking strategies to reconstruct missing regions, effectively capturing deep semantic features and inter-slice anatomical correlations. Similarly, Spark3D [31, 32] systematically explores 3D CNN-based MIM configurations, demonstrating significant gains through large-scale pretraining. Parallel to MIM, contrastive frameworks like VoCo [35,36] exploit consistent spatial anatomy by contrasting arbitrary sub-volumes against fixed "base crops," thereby implicitly encoding high-level positional priors. More recently, works like BrainMVP [26] have extended SSL to cross-modal settings for brain MRI, utilizing cross-sequence reconstruction to align features across different contrasts. Despite these advances, most existing methods treat scans in isolation, focusing strictly on *intra-volume* consistency. Consequently, they fail to exploit the rich structural correspondence inherent in naturally paired multi-phase data (e.g., NCCT-CECT), leaving a critical gap in learning phase-invariant representations.

2.2 Cross-Phase Medical Image Analysis

Clinical workflows frequently acquire multi-phase scans, such as non-contrast (NCCT) and contrast-enhanced CT (CECT), which share consistent anatomy but differ visually due to contrast dynamics. While this has spurred task-specific multi-phase methods, they remain largely confined to supervised settings. For segmentation and diagnosis, frameworks like MULLET [34] and cross-phase attention networks [30] aggregate multi-phase features to refine lesion analysis. Similarly, DuoProto [38] aligns class-specific prototypes across phases for

prognosis, and Bai et al. [2] introduced Cross-Phase Mutual Learning (CPMN) for pulmonary embolism, using inter-feature alignment to transfer knowledge from CTPA to NCCT via dual-phase supervised training. Although cross-modal self-supervised learning (SSL) has shown promise in natural vision (e.g., RGB-Infrared alignment [22]), its adaptation to multi-phase medical pretraining remains underexplored. Crucially, existing medical approaches predominantly rely on *full supervision* or *task-specific constraints* that enforce voxel-level alignment for global semantics. They lack a general, unsupervised pretraining paradigm capable of learning robust, phase-invariant anatomical representations independent of downstream labels.

2.3 Structure- and Topology-Aware Learning

Representations that capture anatomical structure beyond pixel-level appearance improve robustness to domain and appearance shifts. To highlight consistent boundaries, frequency-domain and wavelet-based methods preserve critical spatial information [24, 28]. For instance, extracting multi-scale edges via Haar wavelets [39] or wavelet scattering transforms [18] allows models to downsample without losing high-frequency anatomical features. Similarly, graph- and topology-based models explicitly encode the geometric relationships between anatomical components. Graph neural networks (GNNs) capture complex inter-regional patterns by modeling segmented regions or landmarks as nodes with spatial edges [15]. In a parallel effort to preserve geometry, topology-aware approaches use persistent homology [3] to penalize anatomically implausible segmentation errors, such as spurious holes or disconnected components. While effective, these techniques predominantly operate within a single modality and heavily rely on supervised, task-specific losses (e.g., discrete mask penalties). This leaves the development of a general, unsupervised cross-phase structural representation underexplored.

3 Method

3.1 Overall Framework

To leverage the structurally corresponding yet visually heterogeneous nature of multi-phase volumetric scans, we propose **CAST** (**C**ross-phase **A**lignment of **S**tructure and **T**opology). As illustrated in Fig. 2, given paired unenhanced NCCT volumes X_{ncct} and contrast-enhanced CECT volumes X_{cect}, both inputs are processed by a shared 3D encoder f_θ based on the ResEnc U-Net architecture [12].

The NCCT branch maps its multi-scale encoder features through a lightweight predictor p_ψ, yielding predicted representations for cross-phase alignment, while the CECT branch provides the corresponding structure-amplified target features. Beyond standard volumetric reconstruction, CAST enforces structural alignment between the predicted NCCT features and CECT target features through two complementary feature-level constraints: *Spectral Consistency* and *Topological Consistency*.

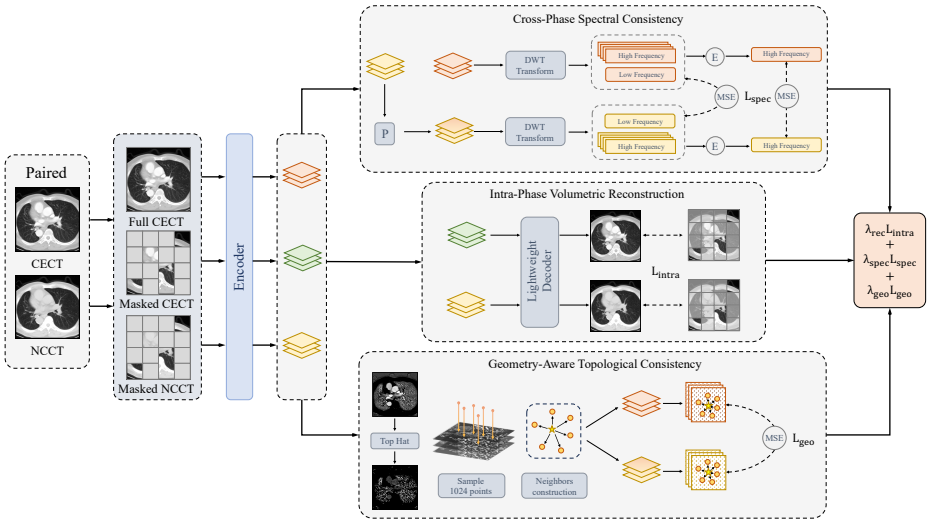

Fig. 2: Overview of the proposed **CAST** framework. Given paired NCCT and CECT volumes, a shared encoder extracts multi-scale features from masked inputs, and a lightweight predictor maps NCCT features to the CECT target feature space for cross-phase alignment. CAST is optimized by three complementary objectives: intra-phase volumetric reconstruction ($\mathcal{L}_{intra}$), cross-phase spectral consistency ($\mathcal{L}_{spec}$) using DWT-based frequency alignment, and geometry-aware topological consistency ($\mathcal{L}_{geo}$) using Top-Hat-mined keypoints and local relational graphs. The final loss combines these terms to learn phase-invariant structural representations.

3.2 Intra-Phase Volumetric Reconstruction

To ensure the 3D CNN encoder acquires foundational volumetric contexts and local spatial continuity, we first apply a standard reconstruction objective [11]. We apply a block-wise masking strategy to both input volumes. The decoder, consisting of five sequential 3D transpose convolution blocks, aims to reconstruct the missing voxel intensities from the bottleneck features.

The intra-phase reconstruction loss L_{intra} is formulated as the average Mean Squared Error (MSE) between the reconstructed outputs $\hat{X}$ and the original unmasked inputs X for both the NCCT and CECT streams:

$$L_{intra} = \frac{1}{2}\left(\text{MSE}(\hat{X}_{NCCT}, X_{NCCT}) + \text{MSE}(\hat{X}_{CECT}, X_{CECT})\right)$$

This foundational objective ensures that the network retains the basic capability to model local appearance within a given phase before complex cross-phase alignment is introduced.

3.3 Cross-Phase Spectral Consistency

Because CECT scans act as structure-amplified views, naively matching raw NCCT and CECT encoder features would force the network to overfit to phase-

specific intensity variations (e.g., the brightened vascular beds). To extract phase-invariant structural boundaries while discarding contrast-induced textural noise, we introduce a Cross-Phase Spectral Consistency constraint utilizing a 3D Discrete Wavelet Transform (DWT) [21].

3D Wavelet Decomposition. Given a feature map $x \in \mathbb{R}^{C\times D\times H\times W}$, the 3D DWT applies a sequence of low-pass (L) and high-pass (H) filters along the depth, height, and width spatial dimensions. This decomposes the representation into one low-frequency volumetric component x_{LLL} and seven high-frequency subbands $\mathcal{H} = \{x_{LLH}, x_{LHL}, x_{LHH}, x_{HLL}, x_{HLH}, x_{HHL}, x_{HHH}\}$. We apply this decomposition to the predicted NCCT features s produced by the lightweight predictor p_ψ and the corresponding CECT target features t from the last three encoder stages.

Low-Frequency Context Alignment. The LLL subband captures the smooth, global anatomical context (e.g., soft tissue distributions and organ topology), which remains largely stable across different contrast phases. Therefore, we explicitly align these base anatomical representations by computing the direct Mean Squared Error (MSE) between s_{LLL} and t_{LLL}.

High-Frequency Energy Alignment. The seven high-frequency subbands encode sharp structural boundaries and localized details. Since the contrast agent drastically amplifies the intensity gradients at these boundaries, directly matching the high-frequency coefficients would compel the NCCT stream to hallucinate the contrast agent's appearance. Instead, we align the *structural energy* of these subbands. For a given feature scale i, we compute the channel-wise high-frequency energy as $E(x_h) = \sqrt{\frac{1}{C}\sum_{c=1}^{C}(x_h^c)^2}$ for each $h \in \mathcal{H}$. This extracts the spatial presence of boundaries independent of their absolute amplitude.

The overall spectral loss L_{spec} across all N aligned scales is formulated as:

$$L_{spec} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{MSE}(s_{LLL}^{(i)}, t_{LLL}^{(i)}) + \lambda_E^{(i)}\sum_{h\in\mathcal{H}}\mathrm{MSE}(E(s_h^{(i)}), E(t_h^{(i)}))\right)$$

where $\lambda_E^{(i)}$ represents scale-specific energy weights. We empirically set $\lambda_E^{(i)}$ to 20.0, 10.0, and 5.0 for the highest to lowest resolution scales, respectively. This hierarchical weighting penalizes structural boundary mismatches more heavily at finer resolutions, where anatomical details are most prominent.

3.4 Geometry-Aware Topological Consistency

While spectral consistency aligns boundary frequencies, it operates globally and independently at each spatial location. However, the true phase-invariant signature of human anatomy lies in the local geometric relationships between adjacent structures. To explicitly preserve these spatial configurations across phases, we propose a Geometry-Aware Topological Consistency module.

Salient Keypoint Mining via Top-Hat. Rather than performing dense voxel-wise alignment, which introduces background noise and computational redundancy, we dynamically mine salient anatomical keypoints (e.g., vascular bifurcations) driven by the contrast agent. Let $D = \mathrm{ReLU}(X_{CECT} - X_{NCCT})$ denote

the non-negative difference map highlighting the enhanced regions. To filter out diffuse background enhancements and isolate sharp structural disparities, we apply a 3D morphological Top-Hat operation $T(\cdot)$ [9]:

$$T(D) = D - (D \circ S)$$

where $\circ$ denotes the morphological opening operation, and S represents a 3D structuring element efficiently implemented via 3D max-pooling. The filtered map $T(D)$ is then normalized into a spatial probability distribution, from which we sample $N_s = 1024$ coordinate points as structural anchors.

Relational Graph Construction. Using grid sampling, we extract node features at these coordinates from the predicted NCCT feature maps and the CECT target feature maps, yielding node sets $\mathcal{V}_{NCCT}$ and $\mathcal{V}_{CECT}$. Although the paired volumes are co-registered to establish anatomical correspondence, coordinate-based feature extraction in the downsampled latent space makes the topological modeling robust to minor residual registration errors. For each phase, we construct a local K-Nearest Neighbors (KNN) relational graph $G = (\mathcal{V}, \mathcal{E})$ based on spatial Euclidean distance, identifying $K = 27$ spatial neighbors for each center node c.

Directional Edge Alignment. For each center node c and its neighbor $n \in \mathcal{N}(c)$, we compute the directed edge vector in the high-dimensional feature space: $v = f_n - f_c$. Since the absolute magnitude of this feature vector is sensitive to contrast-induced intensity shifts, we apply L_2-normalization to isolate the geometric topology and focus on relative directions. The topological loss L_{geo} enforces consistency between the normalized relational graphs across the two phases over a batch size of B:

$$L_{geo} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{N_s K}\sum_{c\in\mathcal{V}}\sum_{n\in\mathcal{N}(c)} \text{MSE}\left(\frac{v_{NCCT}^{c\to n}}{||v_{NCCT}^{c\to n}||_2+\epsilon}, \frac{v_{CECT}^{c\to n}}{||v_{CECT}^{c\to n}||_2+\epsilon}\right)$$

where $\epsilon = 10^{-6}$ ensures numerical stability. By aligning these directional graphs, the network preserves intrinsic structural topology independent of phase-specific appearance.

3.5 Learning and Optimization

Overall Loss. The overall pretraining objective of CAST is a weighted sum of the intra-phase reconstruction loss, the cross-phase spectral consistency loss, and the geometry-aware topological consistency loss. The total loss is defined as:

$$L_{total} = \lambda_{rec}L_{intra} + \lambda_{spec}L_{spec} + \lambda_{geo}L_{geo}$$

where λ_{rec}, λ_{spec}, and λ_{geo} are trade-off hyperparameters controlling the relative importance of each constraint. In our default setting, we empirically set $\lambda_{rec} = 1.0$, $\lambda_{spec} = 1.0$, and $\lambda_{geo} = 1.0$.

Implementation Details. During pretraining, the input patch size is set to $96 \times 224 \times 224$. We utilize a masking block size of 16 and a masking ratio of 75%.

Algorithm 1 CAST Pretraining Forward Pass

Require: $X_{\text{NCCT}}, X_{\text{CECT}}$, f, p_ψ, g, DWT, $\lambda_{\text{rec}}, \lambda_{\text{spec}}, \lambda_{\text{geo}}, \lambda_E^{(i)}$

for $(X_{\text{NCCT}}, X_{\text{CECT}})$ **in** loader:
  # 1. Multi-scale features & reconstruction
  $feat_{\text{NCCT}} \leftarrow f(\text{Mask}(X_{\text{NCCT}}))$; $feat_{\text{CECT}} \leftarrow f(\text{Mask}(X_{\text{CECT}}))$
  $pred_{\text{NCCT}} \leftarrow p_\psi(feat_{\text{NCCT}})$
  $\widehat{X}_{\text{NCCT}}, \widehat{X}_{\text{CECT}} \leftarrow g(feat_{\text{NCCT}}), g(feat_{\text{CECT}})$
  $L_{\text{intra}} \leftarrow \text{MSE}(\widehat{X}_{\text{NCCT}}, X_{\text{NCCT}}) + \text{MSE}(\widehat{X}_{\text{CECT}}, X_{\text{CECT}})$

  # 2. Spectral Consistency via DWT $(L_{\text{spec}} \leftarrow 0)$
  for $i, (s, t)$ **in** enumerate(zip($pred_{\text{NCCT}}, feat_{\text{CECT}}$)):
    $(s_{\text{LLL}}, s_{\text{high}}) \leftarrow \text{DWT}(s)$; $(t_{\text{LLL}}, t_{\text{high}}) \leftarrow \text{DWT}(t)$
    $L_{\text{spec}} \mathrel{+}= \text{MSE}(s_{\text{LLL}}, t_{\text{LLL}})$
      $+\lambda_E^{(i)}\text{MSE}(\text{Energy}(s_{\text{high}}), \text{Energy}(t_{\text{high}}))$

  # 3. Topological Consistency via Top-Hat
  $D \leftarrow \text{TopHat}(\text{ReLU}(X_{\text{CECT}} - X_{\text{NCCT}}))$
  $P \leftarrow \text{SampleKeypoints}(D, 1024)$; $G \leftarrow \text{BuildKNN}(P, 27)$
  $E_{\text{NCCT}}, E_{\text{CECT}} \leftarrow \text{GetEdges}(pred_{\text{NCCT}}, P, G), \text{GetEdges}(feat_{\text{CECT}}, P, G)$
  $L_{\text{geo}} \leftarrow \text{MSE}(\text{L2\_Norm}(E_{\text{NCCT}}), \text{L2\_Norm}(E_{\text{CECT}}))$

  # 4. Overall Objective: $Loss \leftarrow \lambda_{\text{rec}}L_{\text{intra}} + \lambda_{\text{spec}}L_{\text{spec}} + \lambda_{\text{geo}}L_{\text{geo}}$
  `opt.zero_grad(); Loss.backward(); opt.step()`

The model is trained for 1000 epochs with 250 iterations per epoch and a total batch size of 8. Gradients are clipped at a maximum norm of 12, and Automatic Mixed Precision (AMP) is utilized to accelerate the training process.

4 Experiments

To comprehensively evaluate the representations learned by CAST, we conduct extensive experiments on various 3D medical image analysis tasks, as shown in Table 1. Our evaluation aims to answer the following research questions: 1) Can cross-phase alignment improve performance on both NCCT and CECT downstream tasks? 2) Does true patient-level NCCT-CECT pairing provide useful supervision beyond simply using more single-phase CT data? 3) Does CAST exhibit superior label efficiency, faster convergence, and favorable pretraining-scale behavior? 4) How do the individual components and sampling strategies contribute to the overall success?

4.1 Experimental Setup

Datasets and Tasks. We pretrain CAST on **13,850 paired NCCT-CECT scans** covering chest regions. Downstream evaluation spans diverse benchmarks (Table 1): (1) **NCCT tasks** include in-house PE/Coronary segmentation and

Table 1: Summary of datasets and corresponding downstream tasks used for comprehensive validation. The evaluation includes both large-scale real-world clinical cohorts and established public benchmarks to assess generalization across different imaging protocols and anatomical regions.

Category	Dataset	Task
In-house dataset	PE dataset	PE segmentation / classification
	Coronary dataset	Coronary segmentation; Coronary stenosis classification
Public dataset	MSD_Lung [1]	Lung tumor segmentation
	MM-WHS [42]	Heart segmentation
	AIIB23 [23]	Airway segmentation
	Parse22 [19]	Pulmonary vessel segmentation
	SegThor [16]	Organs-at-risk segmentation
	StructSeg19 [27]	Organs-at-risk segmentation
	COVID19-20 [25]	COVID segmentation
	HiPaS [7]	Pulmonary vessel segmentation

public benchmarks for airway (AIIB23), organs-at-risk (SegThor), infection (COVID 19-20), and vessels (HiPaS). (2) **CECT tasks** cover in-house vascular segmentation and public datasets for tumors (MSD_Lung), heart (MM-WHS), vessels (Parse22), and multi-organ structures (StructSeg19). Tasks primarily focus on dense 3D segmentation, with subsets for classification.

Baselines and Metrics. We compare CAST with representative baselines covering **Scratch**, **contrastive learning**, **reconstruction-based/MIM**, and **cross-modal/cross-phase** pretraining. Full baseline details are provided in the supplementary material. We report **Dice (DSC)** and **NSD** for segmentation, and **AUC/Accuracy** for classification.

Implementation Details. CAST and all baselines were implemented with PyTorch and nnssl [32]. Volumes were resampled to $1.5 \times 0.7 \times 0.7$ spacing and cropped to $96 \times 224 \times 224$ patches. Pretraining used a 75% masking ratio, SGD optimizer, batch size 8, 250K steps, and loss weights $\lambda_{rec} = \lambda_{spec} = \lambda_{geo} = 1.0$. For segmentation, a U-Net decoder was fine-tuned with Dice and cross-entropy losses under nnU-Net. For classification, we used 3D global average pooling with an MLP head and fine-tuned with AdamW and binary cross-entropy. Full details are in the supplementary material.

4.2 Main Results on Downstream Tasks

NCCT Tasks. Table 2 presents the results on NCCT downstream tasks. Compared to baselines trained on single-phase data, CAST consistently achieves superior performance across all benchmarks. In boundary-sensitive tasks such as *airway segmentation (AIIB23)* and *organ-at-risk segmentation (SegThor)*, CAST outperforms the strongest baselines (VolumeFusion and SimCLR) by **1.1%** and

Table 2: Segmentation performance on in-house and public NCCT benchmarks. Dice (↑) are reported. Best and second-best results are highlighted with background colors.

Method	In-House		Public Benchmarks				Avg.
	PE-NCCT	Cor-NCCT	AIIB23	SegThor	COVID19-20	HiPaS	
Scratch [12]	0.1722	0.7724	0.8779	0.8982	0.6936	0.8935	0.7180
SimCLR [4]	0.1741	0.7796	0.9013	**0.9028**	0.6774	0.8925	0.7213
MAE [6, 11]	0.1678	0.7629	0.8957	0.8964	0.6271	0.8934	0.7072
SwinUNETR [10]	0.1749	0.7660	0.8949	0.8827	0.7026	0.8881	0.7182
Models Genesis [40]	0.1641	0.7751	0.8994	0.8989	**0.7046**	0.8872	**0.7216**
VolumeFusion [33]	0.1743	**0.7805**	**0.9053**	0.8918	0.6727	0.8883	0.7188
VoCo-v1 [35]	0.1566	0.7567	0.8962	0.8978	0.7012	0.8881	0.7161
VoCo [36]	0.1738	0.7718	0.9012	0.8882	0.6535	0.8868	0.7126
Spark3D [31, 32]	0.1567	0.7745	0.9047	0.8993	0.6913	0.8916	0.7197
BrainMVP [26]	**0.1782**	0.7736	0.8818	0.8995	0.6913	**0.8938**	0.7197
Ours	**0.1758**	**0.7863**	**0.9163**	**0.9035**	**0.7075**	**0.8941**	**0.7306**

Table 3: Segmentation performance on in-house and public CECT benchmarks. Dice (↑) are reported. Best and second-best results are highlighted with background colors.

Method	In-House		Public Benchmarks				Avg.
	PE-CECT	Cor-CECT	MSD_Lung	MM-WHS	Parse22	StructSeg19	
Scratch [12]	**0.5751**	0.8892	0.5519	0.8933	0.7727	**0.8919**	0.7624
SimCLR [4]	0.5671	0.8861	**0.7547**	0.9064	0.7883	0.8899	0.7988
MAE [11]	0.5740	0.8877	0.6005	0.9032	0.7972	0.8894	0.7753
SwinUNETR [10]	0.5477	0.8968	0.6621	0.8848	0.7949	0.8911	0.7796
Models Genesis [40]	0.5472	0.8974	0.6160	0.9021	0.8150	0.8819	0.7766
VolumeFusion [33]	0.5557	0.8874	0.7294	0.8907	0.8078	0.8914	0.7937
VoCo-v1 [35]	0.5471	0.8892	0.7332	0.8992	0.7860	0.8873	0.7903
VoCo [36]	0.5751	0.8864	0.6899	0.8953	0.8160	0.8868	0.7916
Spark3D [31, 32]	0.5696	0.8878	0.7286	**0.9097**	0.8402	0.8906	**0.8044**
BrainMVP [26]	0.5737	**0.8979**	0.6814	0.8953	**0.8457**	0.8822	0.7960
Ours	**0.5787**	**0.8983**	**0.7484**	**0.9105**	**0.8463**	**0.8923**	**0.8124**

0.07% Dice, respectively. On average, our method improves over the second-best performer (Models Genesis) by **1.24%**. Beyond segmentation, CAST also delivers leading performance on classification benchmarks. These results confirm that using CECT as a privileged view during pretraining does not corrupt unenhanced representations; instead, it transfers high-contrast structural priors to the NCCT domain, improving the delineation of ambiguous anatomical boundaries.

CECT Tasks. As shown in Table 3, CAST maintains its competitive edge on CECT downstream tasks, achieving state-of-the-art performance across diverse anatomical structures. It ranks first on key tasks such as *whole heart segmentation (MM-WHS)* and *pulmonary vessel parsing (Parse22)*, with Dice scores of **91.05%** and **84.63%**, respectively. Our method also achieves the highest average performance (**81.24%**), surpassing the strongest baseline (Spark3D) by **0.80%**,

Table 4: Classification performance. Higher is better. Best/second-best are highlighted in red/blue.

Method	PE-NCCT	PE-CECT	Cor-CLS
scratch [12]	0.6358	0.7256	0.7624
MAE [11]	0.6662	0.6988	0.8501
SimCLR [4]	0.6834	0.7318	0.8456
Vol.Fusion [33]	0.6802	0.7190	**0.9179**
VoCo-v1 [35]	0.6813	0.7623	0.9066
Spark3D [31]	**0.6880**	**0.7652**	0.8416
BrainMVP [26]	0.6877	0.7137	0.8518
Ours	**0.6914**	**0.7678**	**0.9363**

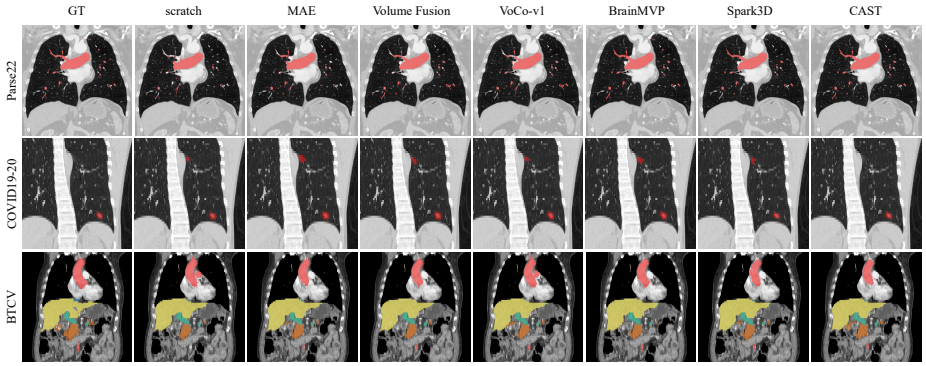

Fig. 3: Qualitative results on Parse22, COVID19-20, and BTCV. Unlike baselines suffering from fragmented vessels, **false positives** in lesions, and blurred boundaries, **CAST** yields continuous structures, clean detections, and sharp contours, validating our spectral and topological constraints.

and yields superior accuracy in CECT-based classification. The results show that CAST is not limited to non-contrast scenarios; cross-phase alignment also refines enhanced-phase representations by preserving phase-invariant topology.

Classification Performance. We further investigate the generalizability of CAST on classification tasks. Table 4 summarizes the performance across three in-house diagnostic cohorts: Pulmonary Embolism (PE) detection on both NCCT and CECT, and Coronary stenosis classification on NCCT. CAST consistently outperforms existing Masked Image Modeling and contrastive baselines. This consistent superiority confirms that our cross-phase pretraining strategy does not merely overfit to local structural boundaries; rather, it successfully aggregates these local topological priors into robust, high-level semantic representations essential for complex clinical diagnostic tasks.

Segmentation Visualizations. Fig. 3 visualizes segmentation predictions across diverse anatomical structures. Compared to conventional MIM baselines, which

often suffer from fragmented vascular branches, lesion false positives, and over-smoothed organ boundaries, CAST predicts more contiguous and sharply defined contours. This visual superiority is evident in fine pulmonary vessels (Parse22), small infections without extra artifacts (COVID19-20), and complex abdominal organs (BTCV), consistent with the benefits of spectral high-frequency alignment and topological consistency.

Dense Feature Similarity via Point Probing. To verify whether CAST learns phase-invariant anatomical representations, we visualize dense feature similarity via point probing in Figure 4. Given query points from representative anatomical regions, we compute their cosine similarity to all spatial locations in the latent feature space. Compared with MAE [11] and VoCo-v1 [35], whose heatmaps are more affected by contrast-induced intensity shifts and irrelevant tissues, CAST produces more coherent and anatomy-aligned similarity patterns. This suggests that CAST better decouples anatomical structure from phase-specific appearance and learns more reliable semantic correspondences.

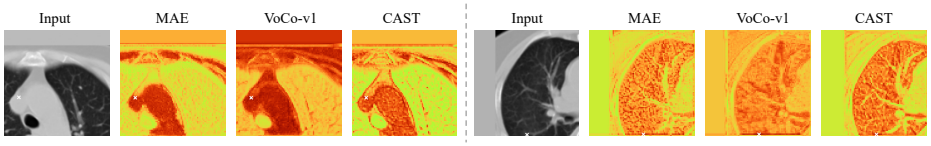

Fig. 4: Feature similarity visualization via point probing. CAST produces more coherent and anatomy-aligned similarity maps than prior pretraining methods.

4.3 Ablation Studies

We conduct comprehensive ablation studies on the **Parse-CECT** and **COVID19-20** datasets to dissect the effectiveness of our proposed modules.

Effectiveness of Core Components. As shown in Table 5(a), the baseline ($\mathcal{L}_{intra}$) achieves 0.7972 Dice on Parse-CTPA. Adding $\mathcal{L}_{spec}$ or $\mathcal{L}_{geo}$ individually yields substantial gains of **4.51%** and **3.93%**, respectively. Crucially, combining both achieves the peak performance (**0.8463**), demonstrating a synergistic effect where global frequency alignment ($\mathcal{L}_{spec}$) and local geometric preservation ($\mathcal{L}_{geo}$) complement each other to learn robust phase-invariant representations.

Impact of Sampling Strategy and Keypoint Density. As shown in Table 5(b), with fixed $N_s = 1024$, our *Diff+TopHat* strategy significantly outperforms *Random* and *Diff* baselines, confirming that top-hat filtering effectively isolates salient anatomical landmarks from background noise. Further ablation on density

Table 5: Ablation studies on core components and sampling strategies. We report Dice scores on Parse-CTPA and COVID19-20. Best and second-best results are highlighted.

(a) Core Components

SPE	GEO	Parse	COVID
		0.7972	0.6271
✓		**0.8423**	0.6935
	✓	0.8365	**0.6975**
✓	✓	**0.8463**	**0.7075**

(b) Sampling Strategy & Density

Strategy	#Pts	Parse	COVID
Random	1024	**0.8268**	0.6986
Diff	1024	0.8021	**0.7054**
D+T	512	0.8174	0.7036
D+T	2048	0.8132	0.7040
D+T	1024	**0.8463**	**0.7075**

Table 6: Comparison of single-phase, shuffled-pair, and paired cross-phase pretraining. Local Sim. and Phase Prob. evaluate anatomical alignment and phase-specific encoding, respectively. Best and second-best downstream results are highlighted.

Method	**COVID19-20**		**MSD_Lung**		**Representation**	
	Dice	NSD	Dice	NSD	Local Sim.↑	Phase Prob.↓
MAE (Standard)	0.6271	0.4403	0.6005	0.5072	0.9608	0.8486
MAE (NCCT-only)	**0.6881**	0.5288	0.6466	0.5249	0.9418	0.8354
MAE (CECT-only)	0.6816	0.5137	**0.7246**	**0.6008**	0.9468	0.8724
CAST (NCCT-only)	0.6878	**0.5422**	0.6796	0.5530	**0.9873**	0.7876
CAST (CECT-only)	0.6675	0.5012	0.7079	0.5569	0.9737	**0.7589**
CAST (Shuffled-Pair)	0.6364	0.4659	0.6676	0.5562	0.9851	0.7664
Ours (Paired Cross-Phase)	**0.7075**	**0.5483**	**0.7484**	**0.6350**	**0.9942**	**0.7146**

($N_s \in \{512, 1024, 2048\}$) reveals that $N_s = 1024$ offers the optimal trade-off: fewer points fail to capture complex topology, while more points introduce noise without performance gains.

Necessity of True Cross-Phase Pairing. To examine whether CAST benefits from true anatomical correspondence rather than merely observing multiple CT phases, we compare standard MAE, single-phase NCCT/CECT pretraining, shuffled-pair CAST, and correctly paired CAST in Table 6. The paired cross-phase setting consistently achieves the best downstream performance, improving Dice to **0.7075** on COVID19-20 and **0.7484** on MSD_Lung. In contrast, shuffling patient-level NCCT-CECT pairs causes clear degradation, confirming that the gain comes from anatomically matched cross-phase supervision rather than additional phase diversity. To further assess representation quality, Local Sim. measures the cosine similarity between paired local NCCT/CECT features, while Phase Prob. evaluates how easily phase information can be decoded from frozen features. Paired CAST achieves the highest Local Sim. and lowest Phase Prob., indicating stronger anatomical alignment and weaker phase-specific encoding.

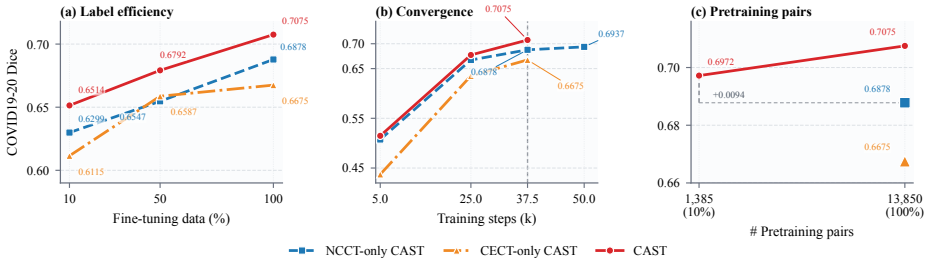

Fig. 5: Label efficiency, convergence, and pretraining-scale analysis on **COVID19-20**. CAST consistently outperforms single-phase variants under limited labels, converges faster during fine-tuning, and benefits from scaling paired pretraining data.

4.4 Label Efficiency, Convergence, and Pretraining-Scale Analysis

We further evaluate CAST on the NCCT **COVID19-20** benchmark under limited annotations, different fine-tuning steps, and different pretraining scales. As shown in Fig. 5(a), CAST consistently outperforms NCCT-only and CECT-only variants across all label budgets. With only **10%** labels, CAST reaches **0.6514** Dice, outperforming NCCT-only and CECT-only CAST by 2.15% and 3.99%, respectively. The advantage remains under 50% and 100% labels, suggesting that paired cross-phase pretraining provides more label-efficient representations. Fig. 5(b) further shows that CAST provides a stronger initialization for downstream optimization. It reaches the best Dice of **0.7075** within **37.5k** fine-tuning steps, whereas NCCT-only CAST remains lower even after extended training to 50k steps. This indicates that the learned cross-phase structural priors not only improve final accuracy but also accelerate adaptation. Finally, Fig. 5(c) studies the effect of pretraining scale. Using only **10%** paired pretraining data, CAST already surpasses the full-scale NCCT-only variant by **0.0094** Dice, and scaling to all 13,850 pairs further improves performance to **0.7075**. These results suggest that true NCCT-CECT correspondence is more valuable than simply increasing single-phase pretraining data, while larger paired datasets can further strengthen phase-invariant representation learning.

5 Conclusion

In this work, we presented CAST, a self-supervised framework that leverages paired NCCT-CECT scans to learn robust, phase-invariant anatomical representations. By introducing Spectral and Geometry-Aware Topological Consistency, CAST effectively decouples structural topology from contrast-induced appearance shifts. Extensive experiments on our large-scale paired dataset demonstrate that CAST achieves state-of-the-art performance across diverse benchmarks while significantly improving label efficiency. This work validates the critical value of cross-phase alignment, offering a powerful paradigm for data-efficient medical image analysis.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Bai, B., Zhou, Y.J., Hu, Y., Mok, T.C., Xiang, Y., Lu, L., Zhang, H., Xu, M.: Cross-phase mutual learning framework for pulmonary embolism identification on non-contrast ct scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 493–503. Springer (2024)
3. Byrne, N., Clough, J.R., Valverde, I., Montana, G., King, A.P.: A persistent homology-based topological loss for cnn-based multiclass segmentation of cmr. IEEE transactions on medical imaging **42**(1), 3–14 (2022)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709 (2020)
5. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9640–9649 (2021)
6. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Balan, M.M., Brown, K.: Masked image modeling advances 3d medical image analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1970–1980 (2023)
7. Chu, Y., Luo, G., Zhou, L., Cao, S., Ma, G., Meng, X., Zhou, J., Yang, C., Xie, D., Mu, D., et al.: Deep learning-driven pulmonary artery and vein segmentation reveals demography-associated vasculature anatomical differences. Nature Communications **16**(1), 2262 (2025)
8. Dao, B.T., Nguyen, T.V., Pham, H.H., Nguyen, H.Q.: Phase recognition in contrast-enhanced ct scans based on deep learning and random sampling. Medical Physics **49**(7), 4518–4528 (2022)
9. Haralick, R.M., Sternberg, S.R., Zhuang, X.: Image analysis using mathematical morphology. IEEE transactions on pattern analysis and machine intelligence (4), 532–550 (1987)
10. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)
12. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
13. Jiang, J., Tyagi, N., Tringale, K., Crane, C., Veeraraghavan, H.: Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (smit). In: International conference on medical image computing and computer-assisted intervention. pp. 556–566. Springer (2022)
14. Jiang, Y., Sun, M., Guo, H., Bai, X., Yan, K., Lu, L., Xu, M.: Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15859–15869 (2023)

15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
16. Lambert, Z., Petitjean, C., Dubray, B., Kuan, S.: Segthor: Segmentation of thoracic organs at risk in ct images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). pp. 1–6. IEEE (2020)
17. Liu, L., Liu, J., Santra, B., Parnell, C., Mukherjee, P., Mathai, T., Zhu, Y., Anand, A., Summers, R.M.: Utilizing domain knowledge to improve the classification of intravenous contrast phase of ct scans. Computerized Medical Imaging and Graphics **119**, 102458 (2025)
18. Liu, R., Nan, H., Zou, Y., Xie, T., Ye, Z.: Lsw-net: A learning scattering wavelet network for brain tumor and retinal image segmentation. Electronics **11**(16), 2616 (2022)
19. Luo, G., Wang, K., Liu, J., Li, S., Liang, X., Li, X., Gan, S., Wang, W., Dong, S., Wang, W., et al.: Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. arXiv preprint arXiv:2304.03708 (2023)
20. Luo, J., Wan, X., Du, J., Liu, L., Zhao, L., Peng, X., Wu, M., Huang, S., Nie, X.: Comprehensive multi-phase 3d contrast-enhanced ct imaging for primary liver cancer. Scientific Data **12**(1), 768 (2025)
21. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE transactions on pattern analysis and machine intelligence **11**(7), 674–693 (2002)
22. Mao, F., Wang, S., Mei, J., Lu, S., Min, C., Liu, F., Feng, X., Wu, M., Hu, Y.: Univ: Unified foundation model for infrared and visible modalities (2025)
23. Nan, Y., Del Ser, J., Tang, Z., Tang, P., Xing, X., Fang, Y., Herrera, F., Pedrycz, W., Walsh, S., Yang, G.: Fuzzy attention neural network to tackle discontinuity in airway segmentation. IEEE transactions on neural networks and learning systems **35**(6), 7391–7404 (2023)
24. Peng, Y., Chen, D.Z., Sonka, M.: Spectral u-net: Enhancing medical image segmentation via spectral decomposition. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
25. Roth, H.R., Xu, Z., Tor-Díez, C., Jacob, R.S., Zember, J., Molto, J., Li, W., Xu, S., Turkbey, B., Turkbey, E., et al.: Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. Medical image analysis **82**, 102605 (2022)
26. Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., Wang, X.: Multi-modal vision pre-training for medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5164–5174 (2025)
27. Shi, J.: Structseg2019 gtv segmentation (2023). `https://doi.org/10.21227/h75x-gt46`
28. Sigillo, L., Grassucci, E., Uncini, A., Comminiello, D.: Generalizing medical image representations via quaternion wavelet networks. Neurocomputing **638**, 130195 (2025)
29. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3 (2025)
30. Uhm, K.H., Jung, S.W., Hong, S.H., Ko, S.J.: Lesion-aware cross-phase attention network for renal tumor subtype classification on multi-phase ct scans. Computers in Biology and Medicine **178**, 108746 (2024)

31. Wald, T., Ulrich, C., Lukyanenko, S., Goncharov, A., Paderno, A., Miller, M., Maerkisch, L., Jaeger, P., Maier-Hein, K.: Revisiting mae pre-training for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5186–5196 (2025)
32. Wald, T., Ulrich, C., Suprijadi, J., Ziegler, S., Nohel, M., Peretzke, R., Kohler, G., Maier-Hein, K.: An openmind for 3d medical vision self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23839–23879 (2025)
33. Wang, G., Wu, J., Luo, X., Liu, X., Li, K., Zhang, S.: Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset (2023)
34. Wu, L., Wang, H., Chen, Y., Zhang, X., Zhang, T., Shen, N., Tao, G., Sun, Z., Ding, Y., Wang, W., et al.: Beyond radiologist-level liver lesion detection on multi-phase contrast-enhanced ct images by deep learning. Iscience **26**(11) (2023)
35. Wu, L., Zhuang, J., Chen, H.: Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22873–22882 (2024)
36. Wu, L., Zhuang, J., Chen, H.: Large-scale 3d medical image pre-training with geometric context priors. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**(3), 3801–3818 (2026)
37. Xing, Z., Zhu, L., Yu, L., Xing, Z., Wan, L.: Hybrid masked image modeling for 3d medical image segmentation. IEEE Journal of Biomedical and Health Informatics **28**(4), 2115–2125 (2024)
38. Yu, H.P., Lyu, S.Q., Hsieh, Y.H., Wang, W., Su, T.H., Kao, J.H., Lin, C.: Learning from limited multi-phase ct: Dual-branch prototype-guided framework for early recurrence prediction in hcc (2025)
39. Zhang, X., Xu, A., Ouyang, G., Xu, Z., Shen, S., Chen, W., Liang, M., Zhang, G., Wei, J., Zhou, X., et al.: Wavelet-guided multi-scale convnext for unsupervised medical image registration. Bioengineering **12**(4), 406 (2025)
40. Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J.: Models genesis. Medical Image Analysis **67**, 101840 (2021)
41. Zhuang, J., Wu, L., Wang, Q., Fei, P., Vardhanabhuti, V., Luo, L., Chen, H.: Mim: Mask in mask self-supervised pre-training for 3d medical image analysis. IEEE Transactions on Medical Imaging (2025)
42. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)