CoInS-Net: A Continuous Position-Aware Network for Joint Medical Image Interpolation and Segmentation

Yujia Sun[†] [a], Ningfeng Que[†] [b], Peiting Shi [b], Rongrong Fu [a], Yingying Yang [a], Xinhang Li [a] and Yin Dai[*] [a]

[a] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, 110169, China
[b] State Key Laboratory of Natural and Biomimetic Drugs, Department of Biomedical Engineering, College of Future Technology, Peking University, Beijing, 100871, China

ABSTRACT

Accurate medical image interpolation and anatomical structure segmentation are fundamental for computer-aided diagnosis and treatment planning. Anisotropic medical volumes with sparse through-plane sampling often suffer from structural discontinuity and boundary blur, hindering reliable clinical image analysis. Most existing methods implement interpolation and segmentation independently, which introduces redundant computation and fails to fully exploit complementary cross-slice structural information between sequential slices. To address these issues, we propose a continuous position-aware interaction network, termed CoInS-Net, for joint frame interpolation and lesion segmentation. Unlike conventional cascaded interpolation-then-segmentation paradigms, the framework enables bidirectional interaction under a shared Swin encoder with continuous spatial coordinate queries. A spatially continuous position interpolation module generates target-position features at every scale from the relative coordinate and physical spacing, and a prototype-based task mutual interaction module lets the segmentation and interpolation branches exchange global structure through a small set of shared prototypes rather than dense feature mixing. A multi-scale task-cooperative decoder further separates each scale into shared and task-specific components, so the two tasks reinforce common anatomy while preserving their distinct requirements down to the boundary level, without extra annotations. Experiments on four public medical imaging datasets with diverse modalities and anatomical regions demonstrate that the proposed method outperforms conventional single-task schemes. It achieves 1.12 to 1.67 dB PSNR gains over the best interpolation baseline and 1.8 to 2.9 percentage point Dice improvements over the best segmentation baseline. The framework also reduces inference latency by nearly half compared with separate single-task pipelines, enhancing the stability and continuity of sequential medical image analysis. The joint optimization framework effectively realizes mutual promotion between interpolation and segmentation tasks, providing a reliable and universal technical scheme for intelligent clinical medical image analysis.

Keywords: Computed Tomography; Magnetic Resonance Imaging; multi-task learning; medical image interpolation; medical image segmentation

1. Introduction

Clinical computed tomography and magnetic resonance imaging volumes are typically highly anisotropic, with substantially higher in-plane resolution than through-plane resolution. Sparse sampling along the z-axis weakens inter-slice anatomical continuity. Organ boundaries, small structures, and lesions often appear blurred, displaced, or abruptly changed across adjacent slices, which increases the difficulty of both intermediate-slice reconstruction and target-position segmentation [1–3].

When the two tasks are modeled independently, interpolation may produce blurred boundaries or anatomically implausible structures due to the lack of semantic constraints [4–8]. Segmentation models cannot fully exploit structural transitions across adjacent slices. The two tasks are inherently complementary. Segmentation semantics can constrain anatomically plausible image synthesis, while inter-slice variations learned through interpolation can facilitate target-structure recognition. Joint modeling of medical slice interpolation and semantic segmentation is therefore highly desirable.

Existing methods that combine medical slice interpolation with semantic segmentation can be broadly divided into two categories. The first category adopts a serial auxiliary strategy. Intermediate images are first generated through frame interpolation, and the densified data are then used for segmentation. In such methods, interpolation serves as a preprocessing step or a form of data augmentation, while the two tasks remain relatively independent. The interpolation process is difficult to optimize according to segmentation requirements, and generation errors may propagate to the segmentation stage [9–12].

The second category integrates interpolation and segmentation into a unified network, enabling joint optimization through cascaded structures, shared encoders, or feature fusion [13, 14]. Although this design strengthens the association between the two tasks, existing methods still mostly rely on generic feature sharing or use interpolation to assist segmentation in a one-way manner. They lack task-specific bidirectional interaction, failing to fully exploit semantic segmentation constraints to guide image generation and to

[†] Yujia Sun and Ningfeng Que contributed equally to this work.
[*] Corresponding author: Yin Dai. E-mail: daiyin@bmie.neu.edu.cn.
ORCID(s):

effectively use cross-slice variations to inform target localization in segmentation. As illustrated in Fig. 1, our method differs from serial auxiliary and conventional joint-learning paradigms by establishing controlled bidirectional interactions between slice interpolation and semantic segmentation.

Simply incorporating interpolation and segmentation into the same network is insufficient to fully exploit their complementarity. The regions where these two tasks most require mutual collaboration are often organ interfaces, lesion boundaries, thin structures, and partial-volume regions with substantial inter-slice variations. In these regions, interpolation requires segmentation semantics to provide anatomical constraints, avoiding excessive boundary smoothing or anatomically implausible structures. Conversely, segmentation requires inter-slice transition information captured during interpolation to enhance structural localization of the target. The key to a joint model therefore lies not merely in feature sharing, but in establishing controllable bidirectional interactions tailored to task-specific requirements.

Such interactive joint modeling still faces several challenges. First, target slice positions are not fixed in anisotropic volumetric data, requiring continuous modeling that incorporates both relative position and physical spacing. Second, the information required by interpolation and segmentation is not completely identical. Simple feature sharing or concatenation may easily introduce negative transfer, so anatomical semantics and inter-slice variation information need to be exchanged selectively. Finally, both tasks rely on boundary details, while additional boundary annotations would reduce the practicality of the method.

To address these issues, we propose CoInS-Net, a continuous position-aware interaction network, for joint medical slice interpolation and semantic segmentation. The network employs a shared Swin Transformer encoder [15] to extract multi-scale features from adjacent slices, and uses two task-specific decoders to simultaneously predict the image and segmentation mask at an arbitrary target position. Specifically, the spatially continuous position interpolation module models the target position and physical spacing to achieve dynamic fusion of adjacent-slice features at every scale. The prototype-based task mutual interaction module exchanges task-specific global information between interpolation and segmentation through a compact set of sample-adaptive prototypes. The multi-scale task-cooperative decoder selectively exchanges spatial information between the two branches through learned gating mechanisms, jointly enhancing image details and segmentation boundaries without requiring additional boundary annotations.

The main contributions of this paper are summarized as follows:

1. We construct a joint framework for arbitrary-position slice interpolation and tumor segmentation in sparse anisotropic medical volumetric data, enabling end-to-end synchronous prediction of intermediate images and their corresponding segmentation masks.
2. We design the spatially continuous position interpolation module, which generates target-position features at every scale from the relative coordinate and physical spacing, with exact endpoint recovery and input-order symmetry built in by construction.
3. We propose the prototype-based task mutual interaction module, which performs bidirectional information exchange between interpolation and segmentation through a compact set of sample-adaptive prototypes, avoiding the unstructured mixing and quadratic cost of dense cross-task attention.
4. We develop the multi-scale task-cooperative decoder, which routes cross-task spatial information through learned sigmoid gates at each upsampling scale, enabling selective feature exchange that refines image details and segmentation boundaries without additional boundary supervision.

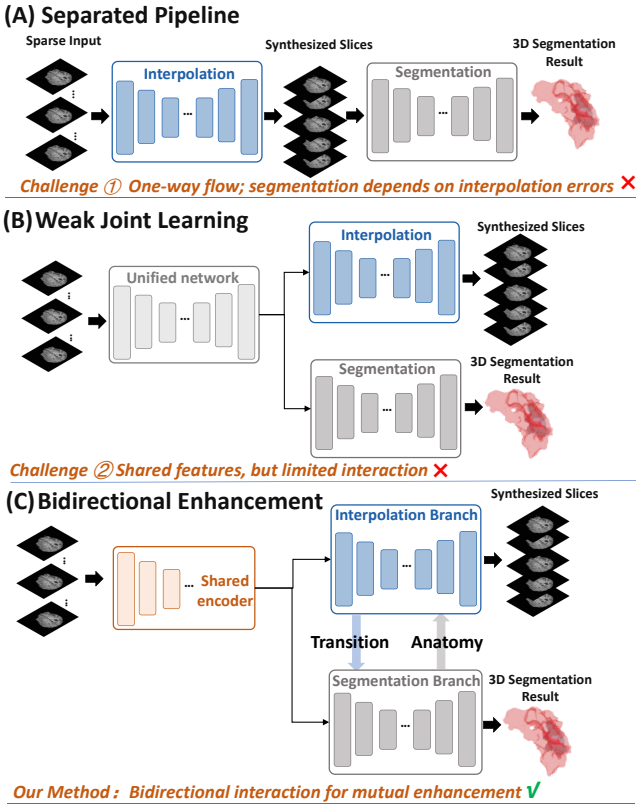


Figure 1: Comparison of three interaction paradigms.

This section reviews prior studies related to slice interpolation for anisotropic medical volumes, volumetric medical image segmentation, and joint learning of slice interpolation and segmentation. The discussion is organized around the three key issues addressed in this work: arbitrary-position prediction, task-specific bidirectional interaction, and collaborative refinement of image details and segmentation boundaries.

1.1. Slice Interpolation for Anisotropic Medical Volumes

To alleviate low through-plane resolution of medical volumetric data, recent slice interpolation methods have gradually evolved from fixed-scale reconstruction [16] toward position- and scale-aware synthesis [4, 17–22]. SAINT incorporates voxel spacing into the slice synthesis process, allowing a single model to accommodate different through-plane upsampling factors [23]. ArSSR represents three-dimensional medical images as continuous implicit voxel functions and enables arbitrary-scale reconstruction through coordinate queries [24]. SA-INR further models local voxel relationships using spatial attention, allowing continuous sampling at arbitrary slice spacings [25]. I^3Net jointly exploits inter-slice variation, in-plane frequency information, and cross-view features to improve recovery of anatomical structures and high-frequency details in anisotropic medical images [26].

These studies demonstrate that arbitrary-scale and continuous-position reconstruction have received considerable attention. Existing methods are primarily designed for image reconstruction. Target coordinates and voxel spacing are generally treated as scale controls, coordinate-query variables, or generation conditions, while the main objective remains image-level reconstruction fidelity. They neither simultaneously predict segmentation results at the target position nor exploit segmentation semantics to constrain the fusion of adjacent-slice features.

The contribution of this work is not merely arbitrary-position slice interpolation. Within a unified interpolation-segmentation framework, we jointly encode the relative target position and physical slice spacing as a continuous position query, which dynamically determines the contributions of the two adjacent slices to the target-position representation at every scale.

1.2. Medical Volumetric Image Segmentation

Medical volumetric image segmentation methods commonly employ three-dimensional convolutional networks [5], Transformer-based architectures [6–8], or 2.5D models to exploit volumetric context. AFTer-UNet separates in-plane feature extraction from through-plane relationship modeling and combines intra-slice and inter-slice contextual information for segmentation of anisotropic medical volumes [27]. Other 2.5D approaches similarly incorporate cross-slice attention or multi-slice feature aggregation to improve utilization of neighboring slices while avoiding the computational cost of fully three-dimensional networks [28–33].

These methods generally assume that the image at the target position has already been observed. Neighboring slices are used to complement the context of the target slice rather than to infer its semantic structure when the target image itself is unavailable. When the target position lies between two sparsely observed slices, organs or lesions may continuously translate, deform, emerge, or disappear along the through-plane direction. Segmentation at such an unobserved position therefore requires more than conventional neighboring-slice aggregation. It also requires explicit modeling of the structural transitions learned during interpolation.

Boundary-aware segmentation constitutes another closely related research direction. Existing methods improve localization of ambiguous structures through auxiliary boundary detection, contour prediction, or signed-distance-map estimation. BA-UNet jointly learns region segmentation and boundary detection and exchanges information between the two tasks [34]. Shape and boundary-aware multi-branch models further employ signed distance maps to emphasize regions near object boundaries [35]. These studies confirm the importance of boundary information for medical image segmentation. They typically introduce explicit boundary-related targets or auxiliary prediction branches. Frequency information and semantic boundaries are usually modeled only within the segmentation task, without exploiting the correspondence between high-frequency image details and semantic boundary responses across interpolation and segmentation.

1.3. Joint Learning of Slice Interpolation and Segmentation

Slice interpolation and semantic segmentation are inherently complementary. Interpolation can recover missing observations in sparsely sampled medical volumes, whereas segmentation semantics can provide anatomical constraints for intermediate-slice synthesis. Existing studies related to these two tasks have mainly adopted serial auxiliary strategies. Slice Imputation generates multiple intermediate images and corresponding labels through frame interpolation, transforms anisotropic volumes into denser representations, and subsequently applies a segmentation network to the augmented data [36]. This approach demonstrates that intermediate-slice synthesis can improve segmentation of anisotropic medical volumes.

Interpolation and segmentation remain sequentially organized in such pipelines. Errors introduced during image synthesis can propagate to the downstream segmentation stage, while segmentation predictions cannot provide feedback to correct anatomically implausible interpolation results. The multi-task learning component [13, 37, 38] in Slice Imputation is used to jointly optimize auxiliary classification and adversarial discrimination objectives within the synthesis model, rather than to establish bidirectional interaction between dedicated interpolation and segmentation decoders.

Joint reconstruction and segmentation frameworks [14, 39–45] commonly employ shared encoders, cascaded prediction, feature concatenation, or separate task heads [46–48]. Although these designs strengthen the connection between image reconstruction and semantic prediction, they do not fully consider the different information requirements of slice interpolation and segmentation. Interpolation primarily focuses on intensity variation, texture transition, and

structural evolution between adjacent slices, whereas segmentation emphasizes category semantics, anatomical localization, and object boundaries. Unstructured mixing of these features may introduce task-irrelevant information and consequently lead to negative transfer.

Existing methods mostly emphasize the use of reconstructed or interpolated images to improve segmentation, while paying limited attention to how segmentation semantics can conversely constrain image synthesis. In anatomically challenging regions such as organ interfaces, lesion boundaries, thin structures, and areas with substantial inter-slice variation, interpolation requires semantic cues to determine whether a generated structure is anatomically plausible. Conversely, segmentation requires transition cues learned by the interpolation branch to estimate the spatial state of a target structure at an intermediate position.

To address these limitations, our method routes the exchange between the two branches through a compact set of sample-adaptive prototypes, so that segmentation semantics and inter-slice transition cues are communicated selectively rather than through direct and unstructured feature sharing. Furthermore, a task-cooperative decoder routes cross-task spatial information through learned gates at each upsampling scale, allowing the model to improve both image structures and segmentation boundaries without introducing an independent boundary-prediction objective or explicit boundary supervision.

2. Method

2.1. Problem Formulation and Architecture Overview

Given two endpoint slices $\boldsymbol{I}_0$ and $\boldsymbol{I}_1$ from the same volumetric scan at physical depths z_0 and z_1, we aim to jointly predict the image $\widehat{\boldsymbol{I}}_t$ and segmentation map $\widehat{\boldsymbol{P}}_t$ at an arbitrary intermediate position parameterized by

$$t = \frac{z_t - z_0}{z_1 - z_0}, \quad t \in (0, 1). \tag{1}$$

Both outputs are produced in a single forward pass, enabling end-to-end mutual optimization between the two tasks.

The proposed network, illustrated in Fig. 2, comprises four stages. A weight-shared Swin Transformer encoder extracts multi-scale features from each endpoint independently (Sec. 2.2). The Physical Position-Aware Feature Generator (PPA-IFG) synthesizes target-position features at every scale using continuous coordinate queries and physical spacing awareness (Sec. 2.3). The Cross-Task Prototype Interaction module (CTPI) establishes bidirectional category-level information exchange between the interpolation and segmentation branches (Sec. 2.5). The Multi-Task Collaborative Decoder (MTC-Decoder) recovers spatial details through dual-branch decoding with a local affinity consistency constraint (Sec. 2.6). The training objective is described in Sec. 2.7.

2.2. Shared Multi-Scale Encoder

We encode each endpoint slice independently through a weight-shared 2D Swin Transformer:

$$\boldsymbol{F}_s^l = \mathcal{E}_l(\boldsymbol{I}_s), \quad s \in \{0, 1\},\ l = 1, \dots, L, \tag{2}$$

where s indexes the endpoint slice and l the feature scale. Weight sharing ensures that corresponding channels across the two endpoints are directly comparable, which is essential for the difference operators and position-conditioned blending used in subsequent modules. Independent encoding also preserves the identity of each endpoint, avoids the positional ambiguity introduced by direct channel concatenation of two slices, and naturally supports input-order swapping during training.

2.3. Physical Position-Aware Feature Generator

The PPA-IFG module generates continuous intermediate features at every encoder scale, conditioned on both the normalized position t and the physical inter-slice spacing. Unlike video frame interpolation methods that operate at fixed temporal intervals, medical volumes exhibit variable slice thickness across patients and protocols. We therefore introduce an explicit physical spacing signal to distinguish cases where the same relative position t corresponds to different amounts of anatomical change.

2.3.1. Physical Position Encoding

We define the normalized spacing ratio

$$\rho = \frac{|z_1 - z_0|}{s_{xy}}, \tag{3}$$

where s_{xy} is the mean in-plane pixel spacing. A position token is then constructed as

$$\boldsymbol{e}_t = \mathrm{MLP}\big[\, t,\ 1{-}t,\ t(1{-}t),\ \log(1{+}\rho),\ \gamma(t) \,\big], \tag{4}$$

with the sinusoidal component $\gamma(t) = [\sin(\pi t), \cos(\pi t), \sin(2\pi t), \cos(2\pi t)]$. We restrict the frequency to low-order harmonics to avoid unstable fitting on limited medical training data. The polynomial term $t(1{-}t)$ encodes the query's distance from both endpoints and peaks at the midpoint where nonlinear correction is most needed.

2.3.2. Linear Position Anchor

At scale l, the linear blend of endpoint features provides a baseline that captures slowly varying anatomy:

$$\boldsymbol{B}_t^l = (1 - t)\, \boldsymbol{F}_0^l + t\, \boldsymbol{F}_1^l. \tag{5}$$

This anchor alone cannot express the nonlinear structural transitions—organ appearance, disappearance, bifurcation, or deformation—that dominate along the through-plane axis of thick-slice acquisitions.

2.3.3. Dual-Endpoint Local Position-Aware Aggregation

To capture nonlinear transitions, we aggregate information from local neighborhoods of both endpoint features,

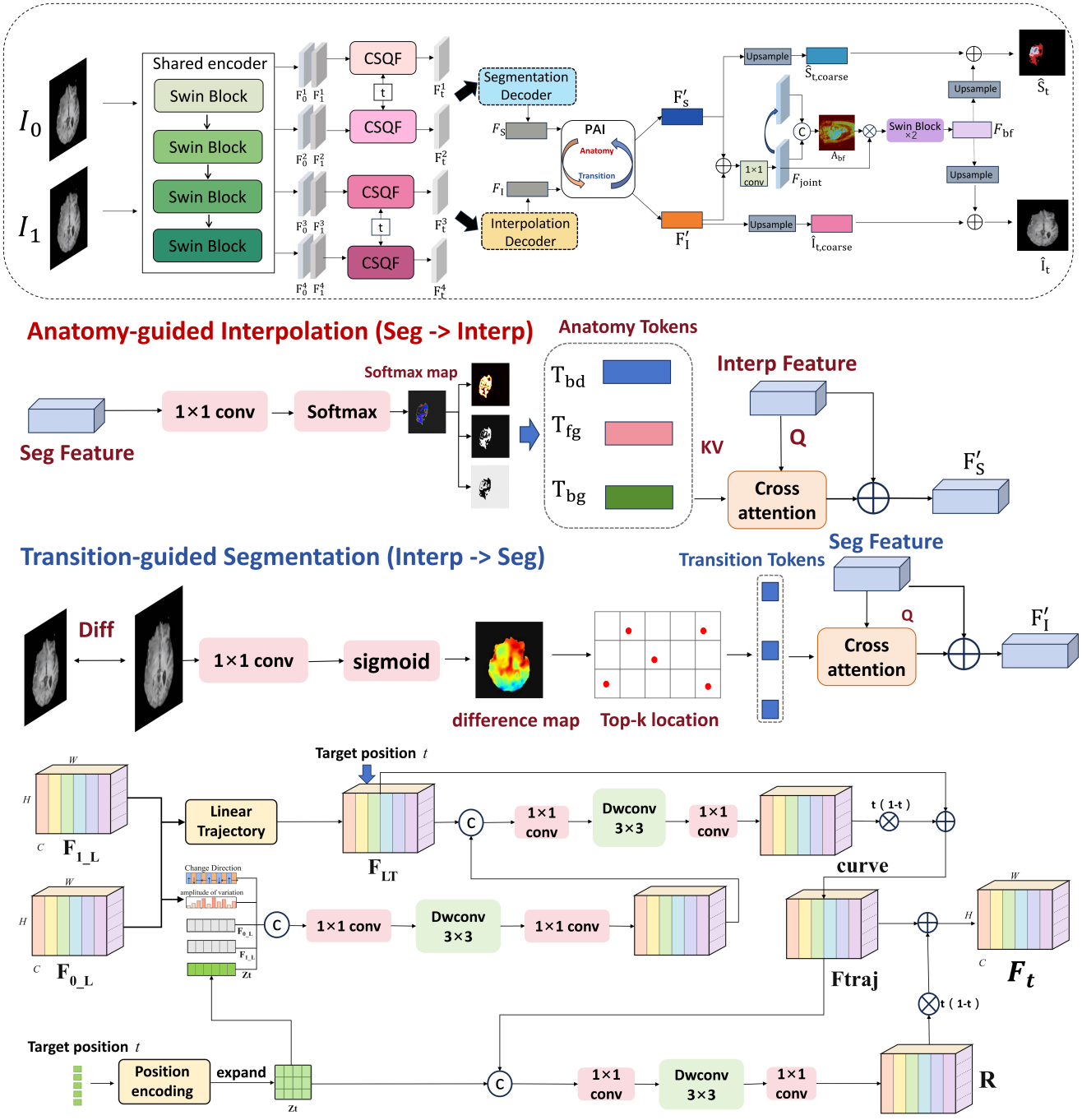


Figure 2: Overall Framework of CoInS-Net

guided by the position token. For each spatial location x, the query is formed from the anchor and position encoding:

$$\boldsymbol{q}_t^l(x) = W_q^l \left[\boldsymbol{B}_t^l(x),\ \boldsymbol{e}_t\right]. \tag{6}$$

Keys and values are extracted from local offsets $\delta \in \mathcal{N}$ of both endpoints $s \in \{0, 1\}$:

$$\boldsymbol{k}_{s,\delta}^l(x) = W_k^l\, \boldsymbol{F}_s^l(x+\delta), \quad \boldsymbol{v}_{s,\delta}^l(x) = W_v^l\, \boldsymbol{F}_s^l(x+\delta). \tag{7}$$

The attention score incorporates a position-dependent bias that favors the closer endpoint:

$$r_{s,\delta}^l(x) = \frac{\boldsymbol{q}_t^l(x)^\top \boldsymbol{k}_{s,\delta}^l(x)}{\sqrt{d}} + \log\big(w_s + \epsilon\big) + b_\delta, \tag{8}$$

where $w_0 = 1-t$, $w_1 = t$, and b_δ is a learnable spatial bias. After softmax normalization over all (s, δ) pairs, the aggregated feature is

$$\boldsymbol{A}_t^l(x) = \sum_{s\in\{0,1\}} \sum_{\delta\in\mathcal{N}} a_{s,\delta}^l(x)\, \boldsymbol{v}_{s,\delta}^l(x). \tag{9}$$

This mechanism jointly encodes position priors, content relevance, and local structural variation, without forcing the medical slice transitions into a 2D optical flow model that is inappropriate for thick-slice acquisitions where structures appear and disappear rather than translate.

2.3.4. Endpoint-Preserving Feature Generation

The final position-aware feature at scale l combines the linear anchor with a gated nonlinear correction:

$$\boldsymbol{P}_t^l = \boldsymbol{B}_t^l + t(1-t)\,\Phi_A^l\left[\boldsymbol{A}_t^l,\ \boldsymbol{B}_t^l,\ \boldsymbol{F}_1^l - \boldsymbol{F}_0^l,\ \boldsymbol{e}_t\right], \tag{10}$$

where Φ_A^l is a lightweight convolutional block. The multiplicative gate $t(1-t)$ guarantees exact endpoint recovery:

$$\boldsymbol{P}_0^l = \boldsymbol{F}_0^l, \qquad \boldsymbol{P}_1^l = \boldsymbol{F}_1^l, \tag{11}$$

so the network learns only the residual that linear blending misses. The gate peaks at t=0.5, where nonlinear correction is most needed, and naturally attenuates near the endpoints where the linear trend dominates.

2.4. Task-Specific Feature Adaptation

The shared position-aware features $\boldsymbol{P}_t^l$ are projected into task-specific subspaces through lightweight adapters:

$$\boldsymbol{U}_I^l = \mathcal{A}_I^l(\boldsymbol{P}_t^l), \qquad \boldsymbol{U}_S^l = \mathcal{A}_S^l(\boldsymbol{P}_t^l), \tag{12}$$

where each adapter consists of a 1×1 convolution, layer normalization, and GELU activation. This separation establishes distinct feature spaces for the interpolation branch ($\boldsymbol{U}_I^l$) and the segmentation branch ($\boldsymbol{U}_S^l$), preventing the two tasks from competing for representational capacity in a single shared space while keeping their inputs aligned through the common position-aware backbone.

2.5. Cross-Task Prototype Interaction

The CTPI module is the core contribution of the proposed framework. It enables bidirectional information exchange between the interpolation and segmentation branches at the category level, using compact prototypes as intermediaries rather than dense pixel-to-pixel attention. The module operates in two steps: cross-task prototype construction and bidirectional prototype-conditioned interaction.

2.5.1. Cross-Task Prototype Construction

We first obtain a coarse semantic probability map from the segmentation features at scale l:

$$\boldsymbol{Q}^l = \mathrm{Softmax}\left(H_{\mathrm{aux}}^l(\boldsymbol{U}_S^l)\right), \tag{13}$$

where $Q_c^l(x)$ indicates the probability of position x belonging to class c. This map serves only to establish spatial category correspondences and is not used as the final segmentation output.

For each class c, we aggregate a segmentation semantic prototype and an interpolation appearance prototype using the same probability weights:

$$\boldsymbol{p}_{S,c}^l = \frac{\sum_x Q_c^l(x)\,\phi_S^l(\boldsymbol{U}_S^l(x))}{\sum_x Q_c^l(x) + \epsilon}, \tag{14}$$

$$\boldsymbol{p}_{I,c}^l = \frac{\sum_x Q_c^l(x)\,\phi_I^l(\boldsymbol{U}_I^l(x))}{\sum_x Q_c^l(x) + \epsilon}, \tag{15}$$

where ϕ_S^l and ϕ_I^l are linear projections. Using the same class weights for both prototypes ensures explicit category-level correspondence: $\boldsymbol{p}_{S,c}^l$ captures the semantic identity and discriminative structure of class c, while $\boldsymbol{p}_{I,c}^l$ captures its intensity distribution, texture, and morphological appearance.

These two prototypes are fused into a cross-task joint prototype:

$$\boldsymbol{m}_c^l = \psi^l\left[\boldsymbol{p}_{I,c}^l,\ \boldsymbol{p}_{S,c}^l,\ \boldsymbol{p}_{I,c}^l \odot \boldsymbol{p}_{S,c}^l\right], \tag{16}$$

where $\odot$ denotes element-wise multiplication that highlights channels where both tasks co-activate, and ψ^l is a two-layer MLP. The concatenation preserves complementary information from both tasks while the element-wise product identifies their shared response patterns.

To avoid noise from absent classes, we restrict subsequent operations to the valid class set $\mathcal{V}^l = \{c \mid \sum_x Q_c^l(x) > \kappa\}$. During early training, the coarse probability is stabilized by mixing with downsampled ground-truth labels: $\widetilde{\boldsymbol{Q}}^l = \eta\,\boldsymbol{Y}_t^l + (1-\eta)\,\mathrm{sg}(\boldsymbol{Q}^l)$, where η anneals from 1 to 0 over the first 20% of training epochs.

2.5.2. Bidirectional Prototype-Conditioned Interaction

The exchange between tasks proceeds in two symmetric directions, each using the other task's prototypes to selectively enhance its own features.

Segmentation-conditioned interpolation. Each spatial location of the interpolation feature queries the segmentation semantic prototypes to determine its category affinity:

$$\alpha_{I,c}^l(x) = \mathrm{Softmax}_{c\in\mathcal{V}^l}\left(\frac{\mathrm{sim}\left(W_{q,I}^l\,\boldsymbol{U}_I^l(x),\ W_{k,S}^l\,\boldsymbol{p}_{S,c}^l\right)}{\tau}\right), \tag{17}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. The aggregated enhancement from the joint prototypes is

$$\boldsymbol{C}_I^l(x) = \sum_{c\in\mathcal{V}^l} \alpha_{I,c}^l(x)\,W_{v,I}^l\,\boldsymbol{m}_c^l, \tag{18}$$

yielding the enhanced interpolation feature:

$$\widetilde{\boldsymbol{U}}_I^l = \Phi_I^l\left[\boldsymbol{U}_I^l,\ \boldsymbol{C}_I^l\right]. \tag{19}$$

This direction provides anatomical boundary awareness and category-level structural priors to the interpolation branch, reducing over-smoothing across tissue interfaces and preventing unrealistic mixing of different anatomical regions.

Interpolation-conditioned segmentation. Symmetrically, the segmentation features query the interpolation appearance prototypes:

$$\alpha^l_{S,c}(x) = \mathrm{Softmax}_{c\in\mathcal{V}^l}\left(\frac{\mathrm{sim}\Big(W^l_{q,S}\,\boldsymbol{U}^l_S(x),\ W^l_{k,I}\,\boldsymbol{p}^l_{I,c}\Big)}{\tau}\right), \tag{20}$$

$$\boldsymbol{C}^l_S(x) = \sum_{c\in\mathcal{V}^l} \alpha^l_{S,c}(x)\, W^l_{v,S}\,\boldsymbol{m}^l_c, \tag{21}$$

$$\widetilde{\boldsymbol{U}}^l_S = \Phi^l_S\big[\boldsymbol{U}^l_S,\ \boldsymbol{C}^l_S\big]. \tag{22}$$

This direction feeds intensity, texture, and morphological transition cues back into the segmentation branch, improving category recognition in low-contrast regions and enhancing boundary localization for small structures and lesions.

The attention weights $\alpha^l_{r,c}(x)$ dynamically select which prototypes are relevant at each spatial position, while the value projections $W^l_{v,r}\,\boldsymbol{m}^l_c$ implicitly determine which channels to transmit. This avoids the quadratic spatial cost of dense cross-attention and provides category-level interpretability: the interaction is mediated by anatomical class prototypes rather than arbitrary spatial correspondences.

2.6. Multi-Task Collaborative Decoder

The CTPI module operates on deep, low-resolution features where semantic abstraction is strongest. The MTC-Decoder complements it by maintaining cross-task cooperation during the high-resolution upsampling stages, where boundary precision and fine structural detail are determined.

2.6.1. Dual-Branch Cooperative Decoding

Let $\boldsymbol{D}^{l+1}_I$ and $\boldsymbol{D}^{l+1}_S$ denote the interpolation and segmentation decoder features at scale $l{+}1$. At each scale, a lightweight shared cooperation feature is first constructed:

$$\boldsymbol{C}^l = \Phi^l_C\big[\mathrm{Up}(\boldsymbol{D}^{l+1}_I),\ \mathrm{Up}(\boldsymbol{D}^{l+1}_S)\big], \tag{23}$$

where $\mathrm{Up}(\cdot)$ denotes bilinear upsampling. Each branch then integrates its own upsampled features, the position-aware skip connection from PPA-IFG, and the shared cooperation signal:

$$\boldsymbol{D}^l_I = \Phi^l_{D,I}\big[\mathrm{Up}(\boldsymbol{D}^{l+1}_I),\ \boldsymbol{P}^l_t,\ \boldsymbol{C}^l\big], \tag{24}$$

$$\boldsymbol{D}^l_S = \Phi^l_{D,S}\big[\mathrm{Up}(\boldsymbol{D}^{l+1}_S),\ \boldsymbol{P}^l_t,\ \boldsymbol{C}^l\big], \tag{25}$$

where $\Phi^l_{D,I}$ and $\Phi^l_{D,S}$ are task-specific convolutional blocks with independent parameters. The skip connections from $\boldsymbol{P}^l_t$ supply position-conditioned multi-scale context, while $\boldsymbol{C}^l$ provides a lightweight communication channel through which the two branches share local structural information without full feature fusion.

The final predictions are obtained from the highest-resolution decoder features:

$$\widehat{\boldsymbol{I}}_t = H_I(\boldsymbol{D}^1_I), \qquad \widehat{\boldsymbol{P}}_t = \mathrm{Softmax}\big(H_S(\boldsymbol{D}^1_S)\big), \tag{26}$$

where H_I and H_S are task-specific output heads.

2.6.2. Spatial Affinity Consistency Constraint

Rather than forcing image gradients to align with segmentation boundaries—which is unreliable in MRI where real anatomical boundaries may have weak intensity contrast and many irrelevant edges exist—we impose a softer constraint: the two branches should agree on local spatial relationships.

From the high-resolution interpolation decoder features $\boldsymbol{D}^1_I$, we compute a normalized structural embedding and define a local affinity:

$$\boldsymbol{Z}_I(x) = \frac{\phi_A(\boldsymbol{D}^1_I(x))}{\|\phi_A(\boldsymbol{D}^1_I(x))\|_2 + \epsilon}, \tag{27}$$

$$A_I(x,q) = \exp\left(-\frac{\|\boldsymbol{Z}_I(x) - \boldsymbol{Z}_I(q)\|^2_2}{\tau_a}\right), \tag{28}$$

where ϕ_A is a linear projection and τ_a is a temperature. This affinity approaches 1 when two positions share similar interpolation structure.

From the segmentation output, we define a complementary affinity based on class probability agreement:

$$A_S(x,q) = \sum_c \widehat{P}_{t,c}(x)\,\widehat{P}_{t,c}(q), \tag{29}$$

which is high when the two positions are predicted to belong to the same class and low at class boundaries.

The spatial consistency loss aligns these two views of local structure over a neighborhood pixel-pair set $\mathcal{E} = \{(x,q) \mid q \in \mathcal{N}(x)\}$:

$$\mathcal{L}_{\mathrm{sc}} = \frac{1}{|\mathcal{E}|}\sum_{(x,q)\in\mathcal{E}} \mathrm{SmoothL1}\big(A_I(x,q) - A_S(x,q)\big). \tag{30}$$

This constraint requires only that the two branches form consistent judgments about which neighboring pixels belong together, without forcing their features to be identical. To prevent background-dominated pixel pairs from overwhelming the loss, we apply balanced sampling across same-class pairs, cross-class pairs, and boundary-adjacent pairs during training.

2.7. Training Objective

The total loss consists of three components corresponding to interpolation reconstruction, segmentation supervision, and cross-task spatial consistency:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{int}} + \lambda_{\mathrm{seg}}\,\mathcal{L}_{\mathrm{seg}} + \lambda_{\mathrm{sc}}\,\mathcal{L}_{\mathrm{sc}}. \tag{31}$$

2.7.1. Interpolation Reconstruction Loss

The interpolation loss combines a robust pixel-level term with a structural similarity term:

$$\mathcal{L}_{\text{int}} = \mathcal{L}_{\text{char}} + \alpha_{\text{ssim}}\,\mathcal{L}_{\text{ssim}}. \tag{32}$$

The Charbonnier loss provides outlier-robust pixel supervision:

$$\mathcal{L}_{\text{char}} = \frac{1}{|\Omega|}\sum_{x\in\Omega}\sqrt{\left(\widehat{I}_t(x) - I_t(x)\right)^2 + \epsilon^2}, \tag{33}$$

while the SSIM loss $\mathcal{L}_{\text{ssim}} = 1 - \text{SSIM}(\widehat{\boldsymbol{I}}_t, \boldsymbol{I}_t)$ preserves local contrast, luminance, and tissue texture continuity. We do not employ adversarial or perceptual losses to avoid hallucinating anatomical structures that do not exist in the ground truth.

2.7.2. Segmentation Loss

The segmentation loss combines Dice and cross-entropy terms to address class imbalance:

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{dice}} + \alpha_{\text{ce}}\,\mathcal{L}_{\text{ce}}, \tag{34}$$

where the Dice loss

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{x,c}\widehat{P}_{t,c}(x)\,Y_{t,c}(x) + \epsilon}{\sum_{x,c}\widehat{P}_{t,c}(x) + \sum_{x,c}Y_{t,c}(x) + \epsilon} \tag{35}$$

provides region-level overlap supervision robust to foreground-background imbalance, and the cross-entropy loss $\mathcal{L}_{\text{ce}} = -\frac{1}{|\Omega|}\sum_x\sum_c Y_{t,c}(x)\log\widehat{P}_{t,c}(x)$ supplies stable per-pixel gradient signals.

2.7.3. Loss Weights and Warm-Up Schedule

The spatial consistency weight follows a linear warm-up schedule:

$$\lambda_{\text{sc}}(e) = \lambda_{\text{sc}}^{\max}\,\min\left(1,\ \frac{e}{E_w}\right), \tag{36}$$

where e is the current epoch and E_w is the warm-up period. This prevents the two branches from imposing erroneous mutual constraints during early training when neither has converged. We set α_{ssim}=0.15, α_{ce}=1, λ_{seg}=1, and $\lambda_{\text{sc}}^{\max}$=0.1 throughout all experiments.

The cross-task prototypes in CTPI are trained end-to-end through the interpolation and segmentation losses without a separate contrastive objective. The category-level alignment emerges naturally from the shared class probability weights in Eqs. (14)–(15) and the downstream task supervision.

2.8. Training Strategy

2.8.1. Training Sample Construction

From a relatively thin-slice or high-resolution reference volume, we select three slices I_i, I_k, I_j with $i < k < j$ and define $\boldsymbol{I}_0 = I_i$, $\boldsymbol{I}_1 = I_j$, $\boldsymbol{I}_t = I_k$ with $t = (z_k - z_i)/(z_j - z_i)$. The corresponding segmentation label is $\boldsymbol{Y}_t = Y_k$. To simulate realistic thick-slice acquisition rather than simple slice deletion, we apply slice-profile blurring along the z-axis before sparse downsampling.

2.8.2. Multi-Position and Multi-Interval Training

We randomly sample varying endpoint intervals (e.g., $j-i \in \{2,3,4,5,6\}$) during training, exposing the network to diverse non-integer relative positions rather than only midpoint interpolation. Input order is randomly swapped as $(\boldsymbol{I}_0, \boldsymbol{I}_1, t) \leftrightarrow (\boldsymbol{I}_1, \boldsymbol{I}_0, 1-t)$ to reinforce the algebraic symmetry built into PPA-IFG and reduce order dependence without requiring an explicit consistency loss.

2.8.3. Two-Stage Optimization

Training proceeds in two phases. In the first phase (approximately 15% of total epochs), CTPI influence is attenuated and the network optimizes $\mathcal{L}_{\text{int}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}$ to establish stable per-task representations. In the second phase, the full CTPI module and spatial consistency loss are activated, and the prototype construction transitions from ground-truth-guided to prediction-guided via the annealing schedule described in Sec. 2.5.

3. Experiments

In this section, we first present the experimental setup including implementation details, datasets, preprocessing and evaluation metrics. We then report quantitative comparative results, provide qualitative visualizations, and evaluate performance under varying sampling spacings and interpolation positions.

3.1. Experimental Setup

3.1.1. Implementation Details

All experiments were conducted on a Linux workstation equipped with an NVIDIA RTX 4090 GPU and 24 GB system memory. The proposed CoInS-Net was implemented in PyTorch and trained with the AdamW optimizer. Mixed-precision training was employed to improve memory efficiency and training stability. The initial learning rate was set to 1×10^{-4} with a cosine annealing scheduler. Training was conducted for up to 100 epochs with early stopping based on validation performance. The model checkpoint achieving the highest combined score of interpolation SSIM and segmentation Dice was retained for final evaluation.

3.1.2. Datasets and Preprocessing

We evaluated CoInS-Net on four public medical image datasets with different modalities, anatomical regions and task complexities to verify the generalization robustness of our multi-task framework. These include the Task02_Heart dataset from Medical Segmentation Decathlon with 21 3D NIfTI volumes for single-class left atrial segmentation on cardiac MRI[1], the Task07_Pancreas dataset from the same benchmark with 282 abdominal CT volumes for dual-class pancreas and tumor segmentation, the AMOS2022 dataset with cross-modal CT and MRI scans and fine-grained annotations of 15 abdominal structures, where the model receives only one modality per case and is required to adapt to both input modalities[2], and the BraTS 2024 post-treatment glioma

[1] http://medicaldecathlon.com/dataaws/

[2] https://amos22.grand-challenge.org/Instructions/

dataset with multi-region pathological MRI annotations, from which the t1n modality is adopted for experiments[3]. All datasets were split at the patient level into training, validation and test with a 7:1:2 ratio.

A consistent data preprocessing pipeline was adopted for all datasets to eliminate potential biases. First, NIfTI files are read via the nibabel library and converted into PyTorch tensors with batch-channel-height-width dimension order. Second, per-image normalization is applied to scale pixel values into the range 0 to 1, eliminating the impact of differences in scanning devices and imaging parameters. Third, data augmentation including random horizontal and vertical flipping, random rotation, and random scaling is applied to the training set to mitigate overfitting and enhance model robustness. For the validation and test sets, only format conversion and normalization are performed.

3.1.3. Evaluation Metrics

Segmentation and interpolation performance are quantitatively assessed using standard metrics. Interpolation quality is measured by Peak Signal-to-Noise Ratio and Structural Similarity Index, while segmentation accuracy is evaluated by Dice Coefficient and Recall.

Peak Signal-to-Noise Ratio is calculated as:

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} (\hat{X}_2(h,w) - X_2(h,w))^2 \\ \mathrm{PSNR} &= 10 \times \log_{10} \left(\frac{\max(X_2)^2}{\mathrm{MSE}} \right) \end{aligned} \tag{37}$$

Here H and W denote frame height and width, X_2 is the ground truth frame, $\hat{X}_2$ is the predicted frame, and $\max(X_2)$ is the maximum pixel value of the ground truth.

Structural Similarity Index is defined as:

$$\begin{aligned} A_{\mathrm{SSIM}} &= (2\mu_{X_2}\mu_{\hat{X}_2} + C_1)(2\sigma_{X_2\hat{X}_2} + C_2) \\ B_{\mathrm{SSIM}} &= (\mu_{X_2}^2 + \mu_{\hat{X}_2}^2 + C_1)(\sigma_{X_2}^2 + \sigma_{\hat{X}_2}^2 + C_2) \\ \mathrm{SSIM}(X_2, \hat{X}_2) &= \frac{A_{\mathrm{SSIM}}}{B_{\mathrm{SSIM}}} \end{aligned} \tag{38}$$

Here μ denotes pixel mean, σ^2 denotes pixel variance, $\sigma_{X_2\hat{X}_2}$ denotes cross-covariance, and C_1, C_2 are regularization constants.

Dice Coefficient is computed as:

$$\begin{aligned} I_i &= \sum_{h=1}^{H} \sum_{w=1}^{W} (\hat{Y}_i(h,w) \times Y_i(h,w)) \\ P_i &= \sum_{h=1}^{H} \sum_{w=1}^{W} \hat{Y}_i(h,w)^2, \quad G_i = \sum_{h=1}^{H} \sum_{w=1}^{W} Y_i(h,w)^2 \\ \mathrm{Dice}(\hat{Y}_i, Y_i) &= \frac{2I_i}{P_i + G_i} \end{aligned} \tag{39}$$

[3] https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation

Here $\hat{Y}_i$ and Y_i are predicted and ground truth segmentation masks respectively.

Recall is defined as:

$$\mathrm{Recall} = \frac{I_i}{G_i} \tag{40}$$

Notations are consistent with the Dice coefficient definition.

Statistical significance was assessed using Welch's t-test across five independent experiments. Single, double and triple asterisks represent $p < 0.05$, $p < 0.01$, and $p < 0.001$ respectively compared with our proposed method. The best performance in each metric column is marked in bold.

3.2. Quantitative Performance Comparison

We evaluate CoInS-Net on four medical image datasets against representative single-task state-of-the-art methods. Interpolation baselines cover residual channel attention architectures, spatial-aware slice synthesis, multi-contrast arbitrary-scale upsampling and implicit neural representation designs, including McASSR [49], SAINT [50], RCAN [51], ACVTT [52], I^3Net [26], CycleINR [53] and SFCLI-Net [54]. Segmentation baselines include convolutional U-shaped networks, transformer backbones, Mamba-based models, diffusion methods and test-time adaptation schemes, namely nnU-Net [5], UNet++ [17], Swin UNETR [8], SegMamba-V2 [55], HiDiff [56], Anatomy-Aware Seg [57] and SicTTA [58].

As shown in Tables 1 and 2, CoInS-Net consistently outperforms all single-task state-of-the-art baselines on both tasks across the four datasets. For interpolation, it yields 1.12 to 1.67 dB PSNR gains over the best baseline SFCLI-Net with universally higher SSIM, verifying that anatomical constraints effectively preserve tissue structure integrity and boundary sharpness. For segmentation, it improves Dice by 1.8 to 2.9 percentage points compared with SegMamba-V2, and the synchronous rise in Recall confirms that inter-slice deformation cues reduce lesion omission. The performance margin widens as dataset difficulty increases, demonstrating the strong generalization of the proposed multi-task interaction framework.

To further investigate the fine-grained segmentation performance of the proposed method on abdominal anatomical structures, we report per-organ segmentation results on the AMOS2022 dataset. We select the five organs with the largest volume fraction, namely the liver, stomach, spleen, left kidney, and right kidney, for detailed comparison. As shown in Table 3, CoInS-Net achieves consistent improvements across all five anatomical structures. The liver with the largest volume and relatively uniform intensity obtains the highest segmentation accuracy and the lowest prediction variance, while the stomach with large morphological variability and irregular cavity contours achieves the lowest performance among the selected organs. The spleen and bilateral kidneys with moderate volume obtain intermediate segmentation accuracy. Stable performance gains across organs of different sizes and morphological complexities demonstrate that the continuous position-aware modeling and prototype interaction mechanism can be generalized to diverse abdominal anatomical structures.

Table 1
Interpolation Performance on Four Medical Image Datasets

Model	MSD Heart		MSD Pancreas		AMOS2022		BraTS 2024	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM	PSNR (dB)	SSIM
RCAN	35.73±0.31***	0.950±0.0041***	34.78±0.29***	0.933±0.0039***	35.15±0.34***	0.930±0.0046***	34.02±0.36***	0.918±0.0049***
SAINT	35.99±0.26***	0.953±0.0034***	35.04±0.24***	0.936±0.0032***	35.42±0.29***	0.935±0.0039***	34.31±0.31***	0.922±0.0042***
McASSR	36.41±0.19***	0.958±0.0026***	35.47±0.18***	0.942±0.0024***	35.80±0.22***	0.941±0.0030***	34.76±0.24***	0.929±0.0033***
ACVTT	35.44±0.35***	0.946±0.0047***	34.57±0.33***	0.930±0.0044***	34.82±0.38***	0.925±0.0051***	33.76±0.40***	0.913±0.0054***
I^3Net	36.11±0.28***	0.954±0.0037***	35.18±0.27***	0.937±0.0036***	35.37±0.31***	0.933±0.0042***	34.45±0.33***	0.924±0.0045***
CycleINR	36.29±0.23***	0.957±0.0031***	35.43±0.21***	0.941±0.0028***	35.69±0.25***	0.938±0.0034***	34.68±0.27***	0.927±0.0037***
SFCLI-Net	36.64±0.16***	0.962±0.0021***	35.70±0.15***	0.945±0.0020***	36.03±0.18***	0.944±0.0024***	35.05±0.20***	0.933±0.0026***
Ours	**37.88±0.21**	**0.975±0.0013**	**37.01±0.08**	**0.958±0.0005**	**37.15±0.14**	**0.958±0.0021**	**36.72±0.16**	**0.949±0.0023**

Table 2
Segmentation Performance on Four Medical Image Datasets

Model	MSD Heart		MSD Pancreas		AMOS2022		BraTS 2024	
	Dice	Recall	Dice	Recall	Dice	Recall	Dice	Recall
UNet++	0.895±0.018***	0.897±0.020***	0.905±0.017***	0.907±0.019***	0.784±0.020***	0.792±0.022***	0.759±0.022***	0.770±0.024***
nnU-Net	0.911±0.014***	0.913±0.016**	0.922±0.013***	0.923±0.015***	0.801±0.016***	0.808±0.018***	0.778±0.018***	0.789±0.020***
Swin UNETR	0.928±0.011***	0.930±0.013***	0.939±0.010***	0.940±0.011***	0.818±0.013***	0.825±0.015***	0.797±0.015***	0.808±0.017***
SegMamba-V2	0.947±0.008***	0.949±0.009**	0.956±0.007**	0.957±0.008***	0.838±0.010***	0.847±0.011**	0.819±0.012*	0.831±0.013***
HiDiff	0.932±0.013**	0.936±0.015***	0.941±0.012***	0.944±0.013***	0.821±0.014***	0.830±0.016**	0.800±0.016*	0.812±0.018***
Anatomy-Aware	0.918±0.016***	0.920±0.018***	0.927±0.015***	0.928±0.017***	0.806±0.018***	0.815±0.020**	0.784±0.020**	0.795±0.022***
SicTTA	0.939±0.010***	0.941±0.011***	0.948±0.009**	0.949±0.010***	0.829±0.011***	0.836±0.013***	0.809±0.013	0.819±0.015***
Ours	**0.968±0.003**	**0.970±0.004**	**0.974±0.006**	**0.975±0.006**	**0.867±0.009**	**0.875±0.010**	**0.842±0.011**	**0.852±0.012**

3.3. Qualitative Visualization

Qualitative comparison on the four datasets is shown in Fig. 3. CoInS-Net generates sharper tissue boundaries in interpolated slices and produces more accurate segmentation masks for target regions, while baseline methods tend to produce blurry edges and miss fine structural details.

3.4. Performance under Different Sampling Spacings

We evaluate joint interpolation and segmentation performance on MSD Pancreas and BraTS 2024 under eight through-plane spacings to verify generalization across clinical scanning protocols. The 4 mm setting matches the default configuration used in prior comparative experiments. Six cascaded interpolation-segmentation model pairs serve as baselines. SSIM and Dice are adopted as core evaluation metrics.

Tables 4, 5, 6, and 7 and Figure 4 present results for different spacing ranges. All methods degrade gradually with increasing spacing due to amplified anatomical deformation and inter-slice discontinuity. CoInS-Net consistently outperforms all cascaded baselines on both datasets across all settings, and its performance advantage widens at larger spacings. This demonstrates that the continuous position-aware interaction mechanism effectively alleviates performance drop under sparse sampling. From 2 mm to 16 mm, Dice decreases by 8.7 percentage points on MSD Pancreas and 9.8 percentage points on BraTS 2024 for our method, versus average declines of 15.8 and 18.4 percentage points for baselines. These results verify that continuous position query modeling stabilizes joint task performance under variable acquisition parameters and exhibits superior robustness compared to cascaded approaches.

We also present sagittal and coronal slice visualizations on subsampled BraTS2024 volumes to comprehensively compare interpolation quality and segmentation accuracy under different z-axis intervals. As shown in Fig. 5, the two-stage baseline still suffers from structural discontinuity

Table 3
Per-organ Segmentation Performance on AMOS2022 Dataset

Model	Liver		Stomach		Spleen		Left Kidney		Right Kidney	
	Dice	Recall	Dice	Recall	Dice	Recall	Dice	Recall	Dice	Recall
UNet++	0.849±0.023***	0.858±0.026***	0.710±0.028***	0.721±0.031***	0.807±0.021***	0.816±0.024***	0.782±0.025***	0.791±0.028***	0.795±0.019***	0.804±0.022***
nnU-Net	0.868±0.017***	0.876±0.020***	0.729±0.022***	0.739±0.025***	0.825±0.019***	0.833±0.022***	0.801±0.021***	0.809±0.024***	0.813±0.016***	0.822±0.018***
Anatomy	0.873±0.020***	0.881±0.023***	0.736±0.025***	0.746±0.028***	0.830±0.023***	0.839±0.026***	0.806±0.018***	0.815±0.021***	0.816±0.020***	0.825±0.023***
UNETR	0.889±0.013***	0.897±0.015***	0.751±0.019***	0.760±0.022***	0.845±0.017***	0.853±0.020***	0.822±0.016***	0.830±0.019***	0.830±0.012***	0.838±0.015***
HiDiff	0.893±0.015***	0.901±0.018***	0.755±0.021***	0.765±0.024**	0.849±0.014***	0.858±0.017***	0.825±0.018***	0.833±0.021***	0.832±0.014***	0.841±0.016***
SicTTA	0.902±0.010***	0.910±0.012***	0.768±0.015***	0.777±0.018**	0.859±0.016**	0.867±0.019**	0.837±0.011***	0.845±0.014***	0.842±0.009***	0.851±0.012***
SegMamba	0.914±0.008***	0.921±0.010**	0.782±0.014**	0.791±0.016**	0.869±0.009***	0.877±0.012**	0.847±0.012**	0.855±0.014**	0.854±0.008***	0.862±0.011***
Ours	**0.937±0.005**	**0.944±0.007**	**0.808±0.009**	**0.818±0.011**	**0.892±0.006**	**0.900±0.008**	**0.870±0.007**	**0.879±0.009**	**0.886±0.006**	**0.895±0.008**

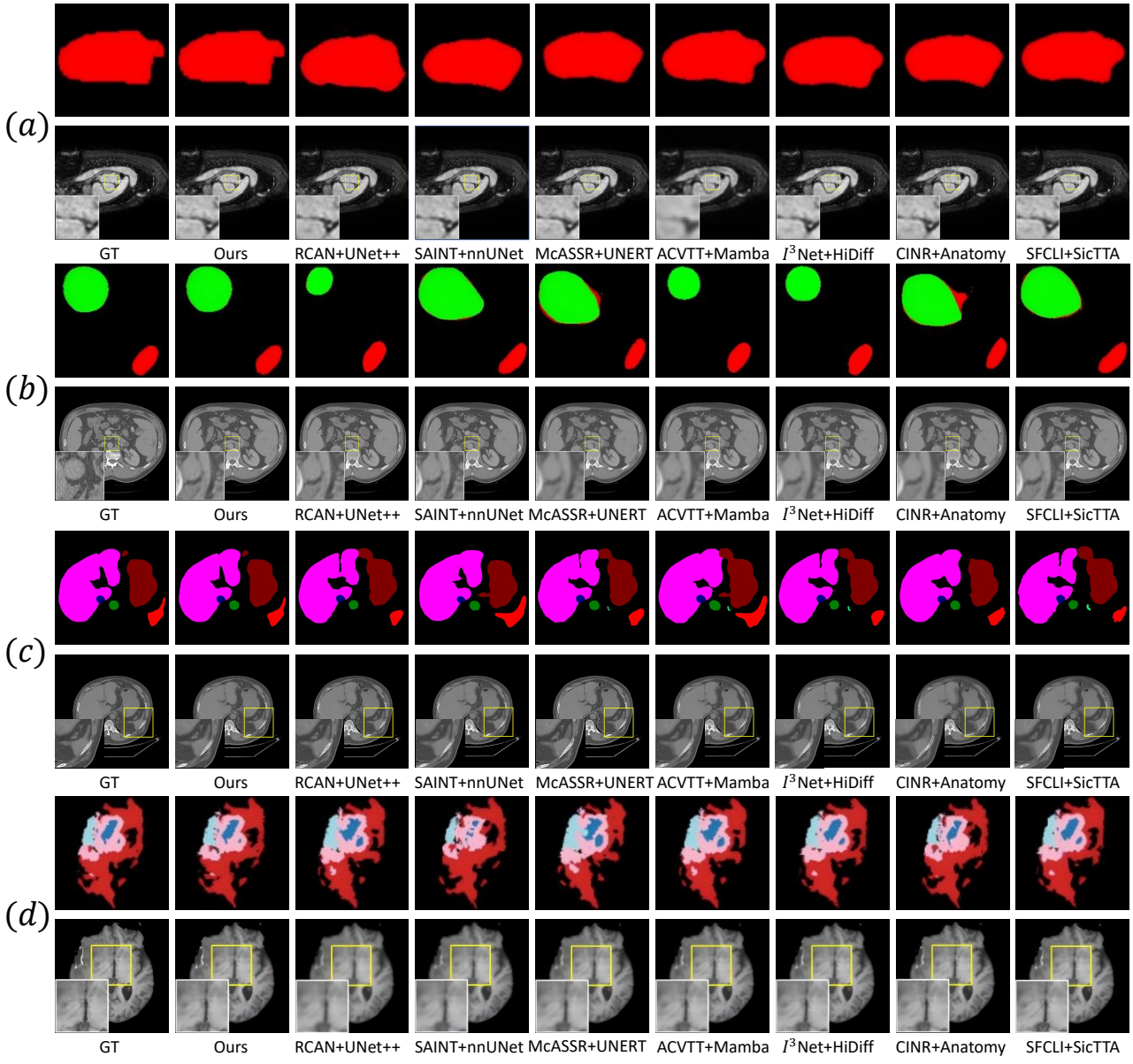


Figure 3: Qualitative results on four datasets. (a) MSD Heart; (b) MSD Pancreas; (c) AMOS2022; (d) BraTS2024.

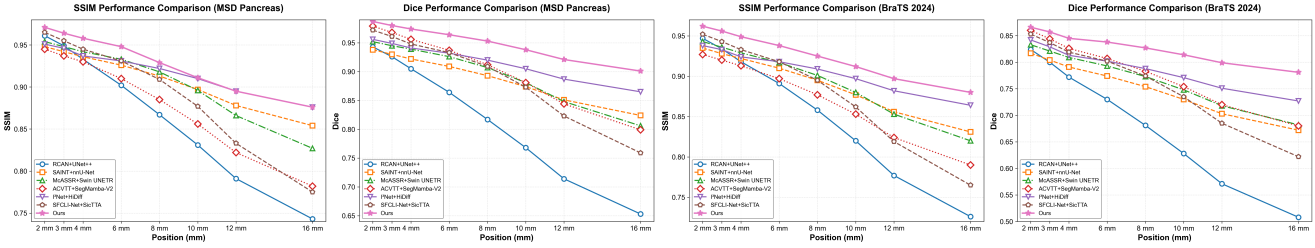


Figure 4: Performance comparison across different z-axis sampling spacings on MSD Pancreas and BraTS 2024 datasets.

and blurry lesion boundaries under large z-axis gaps. Our Position-Aware bidirectional interaction mechanism consistently reconstructs continuous anatomical textures and yields sharp, reliable segmentation masks across all three sampling spacings in both sagittal and coronal views.

3.5. Performance under Different Interpolation Positions

To verify the generalization of the proposed joint framework across arbitrary slice positions, we evaluate interpolation SSIM and segmentation Dice on MSD Pancreas and

Table 4
Performance comparison under 2 mm and 3 mm z-axis sampling spacings on MSD Pancreas and BraTS 2024 datasets

Model	2 mm				3 mm			
	MSD Pancreas		BraTS 2024		MSD Pancreas		BraTS 2024	
	SSIM	Dice	SSIM	Dice	SSIM	Dice	SSIM	Dice
RCAN+UNet++	0.961±0.0028***	0.943±0.014***	0.947±0.0037***	0.824±0.018**	0.949±0.0033***	0.926±0.016***	0.934±0.0042***	0.800±0.019***
SAINT+nnU-Net	0.948±0.0024***	0.938±0.011***	0.935±0.0031***	0.817±0.015***	0.942±0.0028***	0.930±0.012***	0.928±0.0036***	0.804±0.016***
McASSR+Swin UNETR	0.954±0.0018***	0.952±0.008***	0.943±0.0025***	0.833±0.012***	0.948±0.0021***	0.945±0.009***	0.936±0.0029***	0.821±0.013**
ACVTT+SegMamba-V2	0.945±0.0033***	0.979±0.006	0.927±0.0040***	0.860±0.010	0.937±0.0038***	0.968±0.007*	0.920±0.0047***	0.844±0.011**
I^3Net+HiDiff	0.951±0.0027***	0.956±0.010**	0.938±0.0034***	0.842±0.014**	0.946±0.0031***	0.949±0.011**	0.933±0.0039***	0.829±0.015***
SFCLI-Net+SicTTA	0.965±0.0015***	0.972±0.008**	0.952±0.0019***	0.853±0.011***	0.955±0.0018***	0.961±0.009**	0.943±0.0022***	0.837±0.012*
Ours	**0.971±0.0003**	**0.987±0.004**	**0.962±0.0017**	**0.866±0.009**	**0.964±0.0004**	**0.980±0.005**	**0.956±0.0020**	**0.857±0.010**

Table 5
Performance comparison under 4 mm and 6 mm z-axis sampling spacings on MSD Pancreas and BraTS 2024 datasets

Model	4 mm				6 mm			
	MSD Pancreas		BraTS 2024		MSD Pancreas		BraTS 2024	
	SSIM	Dice	SSIM	Dice	SSIM	Dice	SSIM	Dice
RCAN+UNet++	0.933±0.0039***	0.905±0.017***	0.918±0.0049***	0.772±0.020***	0.902±0.0053***	0.864±0.019***	0.891±0.0062***	0.730±0.023***
SAINT+nnU-Net	0.936±0.0032***	0.922±0.013***	0.922±0.0042***	0.791±0.017***	0.926±0.0038***	0.909±0.017***	0.910±0.0056***	0.774±0.015***
McASSR+Swin UNETR	0.942±0.0024***	0.939±0.010***	0.929±0.0033***	0.809±0.014***	0.933±0.0031***	0.926±0.013**	0.917±0.0044***	0.793±0.017***
ACVTT+SegMamba-V2	0.930±0.0044***	0.956±0.007**	0.913±0.0054***	0.826±0.011	0.910±0.0048***	0.937±0.011***	0.897±0.0067***	0.807±0.011*
I^3Net+HiDiff	0.937±0.0036***	0.941±0.012**	0.924±0.0045***	0.813±0.015**	0.931±0.0046***	0.932±0.018*	0.918±0.0051***	0.802±0.019**
SFCLI-Net+SicTTA	0.945±0.0020***	0.948±0.009***	0.933±0.0026***	0.819±0.012*	0.931±0.0034***	0.933±0.014***	0.918±0.0048***	0.802±0.017**
Ours	**0.958±0.0005**	**0.974±0.006**	**0.949±0.0023**	**0.845±0.010**	**0.948±0.0011**	**0.964±0.008**	**0.938±0.0026**	**0.838±0.010**

AMOS 2022 at four relative target positions. All experiments are conducted under a fixed 4 mm through-plane sampling spacing. Three cascaded interpolation-segmentation baselines are adopted for comparison, where position 0.5 corresponds to the default mid-slice setting.

All methods undergo gradual performance degradation as the target position deviates from the slice center, owing to increased anatomical uncertainty. CoInS-Net consistently outperforms all cascaded baselines across all positions on

Table 6
Performance comparison under 8 mm and 10 mm z-axis sampling spacings on MSD Pancreas and BraTS 2024 datasets

Model	8 mm				10 mm			
	MSD Pancreas		BraTS 2024		MSD Pancreas		BraTS 2024	
	SSIM	Dice	SSIM	Dice	SSIM	Dice	SSIM	Dice
RCAN+UNet++	0.867±0.0048***	0.817±0.023***	0.858±0.0061***	0.681±0.021***	0.831±0.0061***	0.768±0.028***	0.820±0.0067***	0.628±0.026***
SAINT+nnU-Net	0.913±0.0065**	0.893±0.018***	0.895±0.0059***	0.754±0.018***	0.900±0.0053**	0.874±0.021***	0.877±0.0063***	0.730±0.020***
McASSR+Swin UNETR	0.918±0.0039*	0.907±0.020**	0.901±0.0054***	0.773±0.016***	0.896±0.0048***	0.881±0.016***	0.880±0.0062***	0.748±0.022***
ACVTT+SegMamba-V2	0.885±0.0071***	0.912±0.012***	0.877±0.0067***	0.783±0.013**	0.856±0.0075***	0.881±0.019***	0.853±0.0074***	0.754±0.015***
I^3Net+HiDiff	0.922±0.0058*	0.920±0.027	0.909±0.0061***	0.788±0.017***	0.910±0.0059	0.905±0.018***	0.897±0.0066***	0.771±0.021*
SFCLI-Net+SicTTA	0.909±0.0051***	0.909±0.015***	0.895±0.0057***	0.774±0.019***	0.877±0.0064***	0.873±0.024***	0.862±0.0069***	0.735±0.023**
Ours	**0.929±0.0021**	**0.953±0.009**	**0.925±0.0031**	**0.827±0.011**	**0.911±0.0028**	**0.938±0.011**	**0.912±0.0037**	**0.814±0.013**

Table 7
Performance comparison under 12 mm and 16 mm z-axis sampling spacings on MSD Pancreas and BraTS 2024 datasets

Model	12 mm				16 mm			
	MSD Pancreas		BraTS 2024		MSD Pancreas		BraTS 2024	
	SSIM	Dice	SSIM	Dice	SSIM	Dice	SSIM	Dice
RCAN+UNet++	0.791±0.0069***	0.714±0.031***	0.777±0.0078***	0.571±0.029***	0.743±0.0076***	0.653±0.035***	0.726±0.0085***	0.508±0.032***
SAINT+nnU-Net	0.882±0.0073*	0.851±0.025***	0.856±0.0071***	0.703±0.024***	0.859±0.0079***	0.824±0.029***	0.831±0.0082***	0.672±0.027***
McASSR+Swin UNETR	0.866±0.0058***	0.848±0.022**	0.853±0.0069***	0.718±0.020***	0.827±0.0065***	0.806±0.026***	0.820±0.0076***	0.682±0.025***
ACVTT+SegMamba-V2	0.822±0.0083***	0.844±0.022**	0.824±0.0087***	0.720±0.018***	0.782±0.0091***	0.799±0.026***	0.790±0.0095***	0.680±0.021***
I^3Net+HiDiff	0.895±0.0068	0.887±0.029*	0.882±0.0075***	0.751±0.024***	0.876±0.0074	0.865±0.033	0.864±0.0083	0.727±0.028*
SFCLI-Net+SicTTA	0.833±0.0074***	0.823±0.028***	0.819±0.0081***	0.685±0.026***	0.775±0.0082***	0.759±0.032***	0.765±0.0089***	0.622±0.029***
Ours	**0.895±0.0035**	**0.921±0.013**	**0.897±0.0044**	**0.799±0.015**	**0.876±0.0042**	**0.901±0.015**	**0.880±0.0051**	**0.781±0.017**

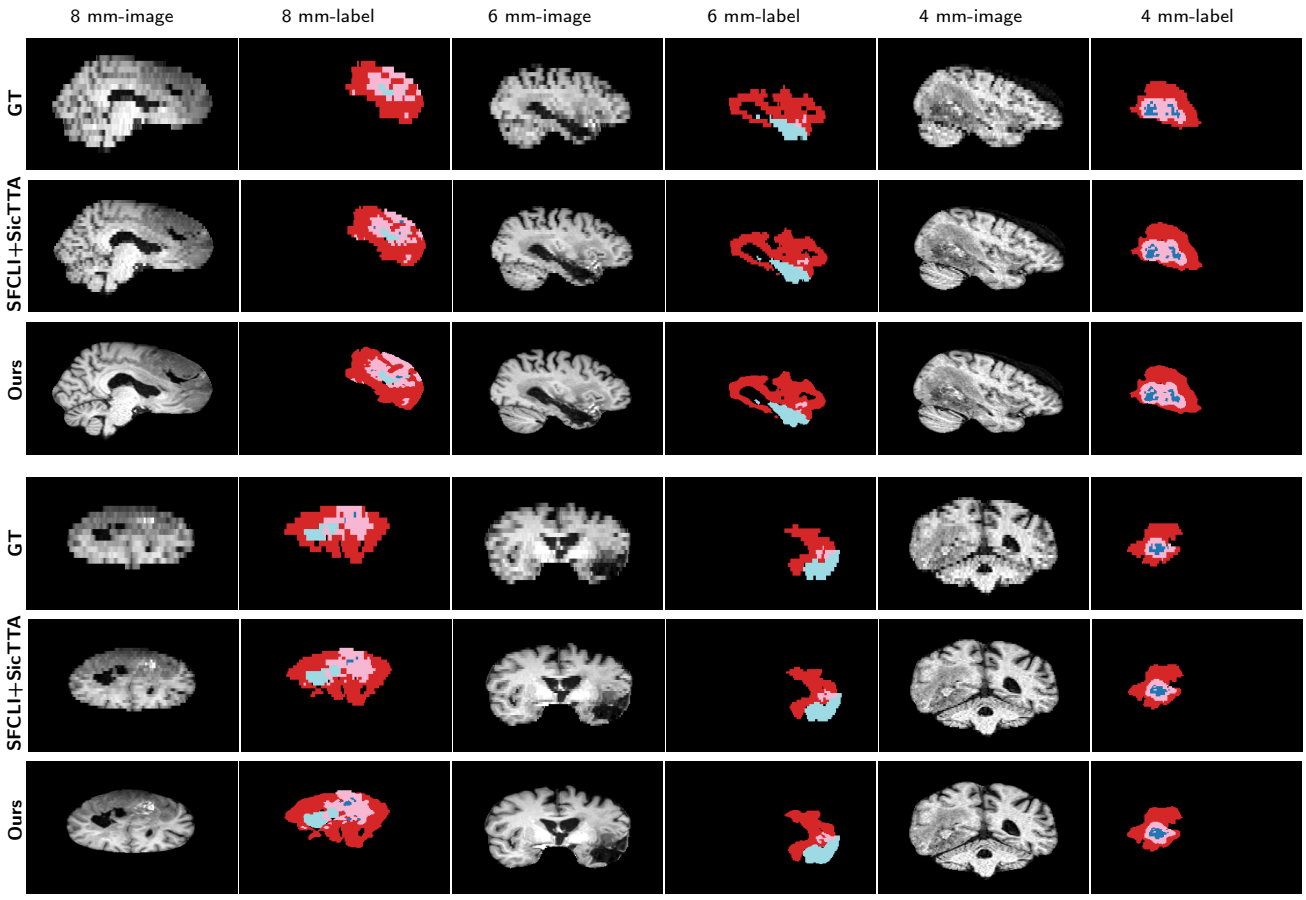


Figure 5: Qualitative results on subsampled BraTS2024 volumes under three z-axis sampling intervals. Rows 1-3 show sagittal views and rows 4-6 show coronal views.

Table 8
Performance comparison at interpolation positions 0.2 and 0.33

Model	Position 0.2				Position 0.33			
	MSD Pancreas		AMOS 2022		MSD Pancreas		AMOS 2022	
	SSIM	Dice	SSIM	Dice	SSIM	Dice	SSIM	Dice
McASSR+Swin UNETR	0.924±0.0038***	0.908±0.016***	0.920±0.0045***	0.779±0.021***	0.940±0.0029***	0.937±0.012***	0.939±0.0034***	0.816±0.015***
I^3Net+HiDiff	0.919±0.0052***	0.910±0.019***	0.912±0.0061***	0.782±0.024**	0.935±0.0033***	0.939±0.014***	0.930±0.0041***	0.819±0.018***
SFCLI-Net+SicTTA	0.920±0.0044***	0.917±0.017**	0.917±0.0053***	0.790±0.022***	0.942±0.0025***	0.946±0.011**	0.936±0.0031***	0.827±0.013**
Ours	**0.946±0.0011**	**0.956±0.009**	**0.946±0.0019**	**0.843±0.012**	**0.956±0.0008**	**0.972±0.008**	**0.956±0.0014**	**0.865±0.010**

both datasets, with a milder performance drop at edge positions. The stable performance benefits from the spatially continuous position interpolation module, which dynamically adjusts feature fusion weights according to relative target coordinates and physical spacing. Quantitative results are presented in Tables 8 and 9.

Qualitative visualization of joint interpolation and segmentation at different positions is shown in Fig. 6. Compared with cascaded baselines, CoInS-Net preserves sharper anatomical boundaries and more accurate segmentation contours at non-central positions, effectively alleviating structural distortion and label misalignment when the target slice deviates from the midpoint.

4. Ablation study

To quantitatively verify the necessity and individual contribution of each core module, we conducted leave-one-out ablation studies, where each variant removed one module from the full model while keeping the same backbone architecture and training hyperparameters for fair comparison. As summarized in Table 10, removing any single component leads to consistent performance degradation across all datasets and both tasks, confirming the indispensability of the proposed design. We ablate the spatially continuous position interpolation module (SCPI), the two directions of the prototype-based task mutual interaction (segmentation-to-interpolation, $\text{PTMI}_{\text{S}\to\text{R}}$, and interpolation-to-segmentation, $\text{PTMI}_{\text{R}\to\text{S}}$), and the multi-scale task-cooperative decoder

Table 9
Performance comparison at interpolation positions 0.5 and 0.66

Model	Position 0.5				Position 0.66			
	MSD Pancreas		AMOS 2022		MSD Pancreas		AMOS 2022	
	SSIM	Dice	SSIM	Dice	SSIM	Dice	SSIM	Dice
McASSR+Swin UNETR	0.942±0.0024***	0.939±0.010***	0.941±0.0030***	0.818±0.013***	0.938±0.0032***	0.935±0.013***	0.937±0.0039***	0.814±0.017***
I^3Net+HiDiff	0.937±0.0036***	0.941±0.012**	0.933±0.0042***	0.821±0.014***	0.933±0.0027***	0.937±0.011***	0.929±0.0035***	0.817±0.014**
SFCLI-Net+SicTTA	0.945±0.0020***	0.948±0.009***	0.944±0.0024***	0.829±0.011***	0.937±0.0041***	0.944±0.015***	0.934±0.0049***	0.825±0.019*
Ours	**0.958±0.0005**	**0.974±0.006**	**0.958±0.0021**	**0.867±0.009**	**0.955±0.0009**	**0.970±0.007**	**0.955±0.0016**	**0.863±0.010**

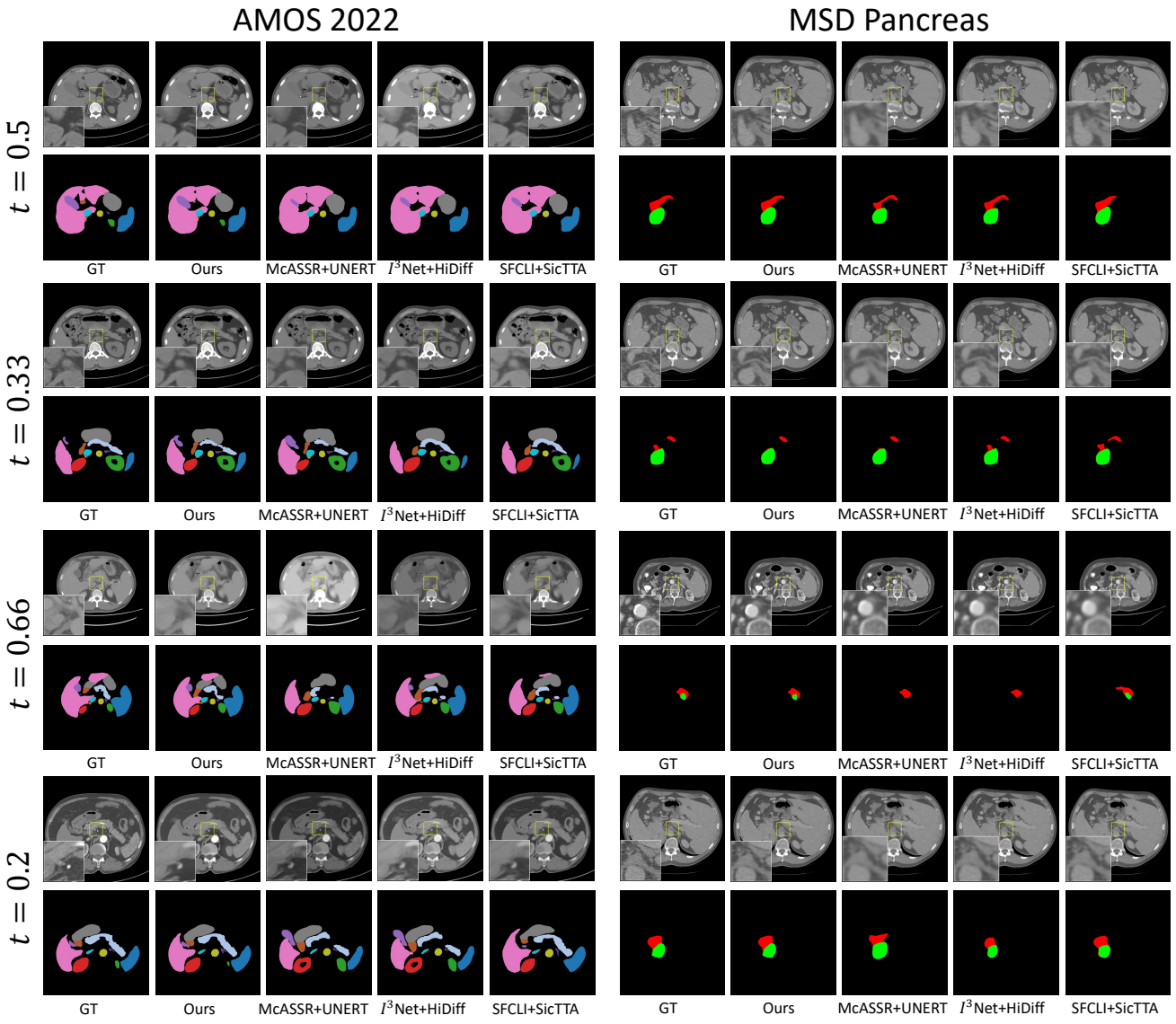


Figure 6: Qualitative comparison at different interpolation positions. Left three columns show interpolation results; right three columns show segmentation results.

(MTC). The segmentation-to-interpolation prototype path contributes most to interpolation, since its removal causes the largest SSIM drop by withdrawing the anatomical structure that keeps synthesised boundaries plausible. For segmentation, the task-cooperative decoder is particularly important, as removing it produces the largest Dice reduction due to degraded fine-grained boundary modeling. Overall, the full model achieves the best performance on all metrics, demonstrating that the three modules jointly provide complementary benefits for the proposed framework.

Table 10
Ablation Studies of Core Modules on Four Medical Image Datasets

Method	MSD Heart		MSD Pancreas		AMOS2022		BraTS 2024	
	SSIM	Dice	SSIM	Dice	SSIM	Dice	SSIM	Dice
w/o SCPI	0.968±0.0031***	0.963±0.014	0.951±0.0016***	0.969±0.018	0.951±0.0038**	0.859±0.022	0.941±0.0046**	0.834±0.025
w/o $\text{PTMI}_{S\to R}$	0.966±0.0044**	0.964±0.020	0.949±0.0021***	0.970±0.024	0.949±0.0052***	0.861±0.031	0.939±0.0063*	0.836±0.034
w/o $\text{PTMI}_{R\to S}$	0.969±0.0026*	0.962±0.011	0.952±0.0011***	0.968±0.015	0.952±0.0030***	0.858±0.019**	0.942±0.0037***	0.833±0.023
w/o MTC	0.970±0.0037	0.959±0.017	0.953±0.0014***	0.965±0.021	0.953±0.0045	0.854±0.027	0.943±0.0055*	0.829±0.030
Ours	**0.975±0.0013**	**0.968±0.003**	**0.958±0.0005**	**0.974±0.006**	**0.958±0.0021**	**0.867±0.009**	**0.949±0.0023**	**0.842±0.011**

Table 11
Comparison of Multi-Task Learning Architectures

Architecture	PSNR (dB)	SSIM	Dice	Recall
Direct Concat	34.38±0.34***	0.933±0.0045***	0.907±0.019***	0.910±0.020***
Shared-Bottom	35.71±0.26***	0.949±0.0035***	0.934±0.015***	0.937±0.016**
Cross-Stitch	35.24±0.30***	0.944±0.0040***	0.923±0.017**	0.926±0.018***
MoE	35.89±0.23***	0.951±0.0031***	0.937±0.013**	0.940±0.014***
PLE	36.21±0.18***	0.955±0.0024***	0.941±0.010***	0.944±0.011**
Ours	**37.88±0.21**	**0.975±0.0013**	**0.968±0.003**	**0.970±0.004**

Table 12
Comparison of Loss Balancing Strategies

Strategy	PSNR (dB)	SSIM	Dice	Recall
Fixed Weight	37.37±0.23**	0.970±0.0031**	0.950±0.013	0.953±0.014***
GradNorm	37.63±0.17*	0.972±0.0023*	0.956±0.010*	0.959±0.011*
DWA	37.69±0.14	0.973±0.0019*	0.959±0.008**	0.962±0.009
Ours	**37.88±0.21**	**0.975±0.0013**	**0.968±0.003**	**0.970±0.004**

4.1. Multi-Task Learning Architecture Comparison

To further validate the superiority of our multi-task design paradigm, we compared five classic multi-task learning architectures on the MSD Heart dataset. All variants shared the same Swin Transformer backbone and training hyperparameters for fair comparison.

Five representative multi-task learning schemes were compared, including Direct Feature Concatenation[37], Shared-Bottom[38], Cross-Stitch[46], Mixture of Experts[47], and Progressive Layered Extraction[48]. Quantitative results are shown in Table 11.

Among all compared architectures, Progressive Layered Extraction achieves the best performance among classic methods with 36.21 dB PSNR and 0.941 Dice, benefiting from its layered expert extraction mechanism that alleviates task conflict. Our CoInS-Net further outperforms this scheme by 1.67 dB in PSNR and 2.7 percentage points in Dice, which is attributed to the fine-grained position modulation and bidirectional prototype-based interaction mechanisms that fully exploit the intrinsic correlation between interpolation and segmentation tasks.

4.2. Loss Balancing Strategy Comparison

We further compared four loss balancing strategies: fixed weighted loss, gradient normalization[59], dynamic weight average[60], and our learnable uncertainty weighting in Eq. (31). Quantitative results are listed in Table 12.

Gradient normalization and dynamic weight average achieve slightly better interpolation performance than the fixed weight strategy, as they dynamically adjust weights according to gradient magnitude or loss decay rate. Their segmentation performance improvement is limited. In contrast, our uncertainty weighting achieves the best comprehensive performance, especially boosting segmentation Dice by 1.8 percentage points compared with fixed weight loss, while maintaining competitive interpolation accuracy. This indicates that adapting the task weights to each task's noise level effectively balances heterogeneous objectives and avoids the model being dominated by the easier interpolation task during training.

5. Model complexity

To further verify the practical deployment value of the proposed CoInS-Net, we conducted comprehensive computational efficiency experiments on the MSD Heart dataset with a batch size of 4 and input size of 512 × 512. We compared with six groups of combined models formed by pairing state-of-the-art interpolation and segmentation models. Quantitative results are summarized in Table 13.

CoInS-Net achieves an average inference time of 37.70 milliseconds per frame, reducing inference latency by 48.1 percent compared with the best combined model. The average inference frame rate reaches 53.09 FPS, meeting the real-time processing requirements of clinical medical image analysis. In terms of model complexity, our model has only 42.98M parameters, 118.23G FLOPs, and 163.96MB model size, all of which are significantly lower than all combined models. For memory consumption during inference, CoInS-Net only occupies 408.92 MB of GPU memory, which is 33.2 percent lower than the best combined model. The significant computational efficiency advantage benefits from the shared encoder architecture that avoids redundant feature extraction, as well as the sparse token interaction mechanism that reduces unnecessary computational overhead, making our model more suitable for deployment in clinical environments with limited hardware resources.

6. Discussion

This study introduces CoInS-Net, an end-to-end single-stage multi-task network for joint medical image interpolation and segmentation. The model achieves state-of-the-art

Table 13
Computational Efficiency Comparison on MSD Heart MRI

Model	Avg Infer Time (ms)	Avg FPS	Parameters (M)	Memory (MB)	FLOPs (G)	Model Size (MB)
RCAN+UNet++	117.83±1.47	16.98±0.23	162.74±0.00	1396.54±14.73	386.92±0.00	1248.37±0.00
SAINT+nnU-Net	108.46±1.13	18.44±0.19	148.39±0.00	1248.76±12.58	342.67±0.00	1095.82±0.00
McASSR+Swin UNETR	96.72±0.91	20.68±0.17	134.86±0.00	1086.43±10.92	304.58±0.00	918.64±0.00
ACVTT+SegMamba-V2	88.35±0.74	22.64±0.15	119.57±0.00	924.71±8.67	268.34±0.00	736.29±0.00
I^3Net+HiDiff	79.28±0.63	25.23±0.13	103.42±0.00	746.38±7.21	231.76±0.00	512.53±0.00
CycleINR+SicTTA	72.64±0.52	27.53±0.11	91.85±0.00	612.47±5.86	189.43±0.00	386.71±0.00
Ours	**37.70±0.22**	**53.09±0.28**	**42.98±0.00**	**408.92±4.61**	**118.23±0.00**	**163.96±0.00**

performance on four mainstream medical imaging datasets covering cardiac MRI, abdominal CT, cross-modal abdominal CT/MRI, and intracranial glioma MRI modalities. The proposed framework fully exploits the intrinsic complementarity between interpolation and segmentation tasks via a shared encoder-dual decoder architecture, combined with continuous position interpolation and bidirectional prototype-based interaction mechanisms, achieving significant performance gains over both single-task baselines and classic multi-task learning architectures.

The core technical innovations of CoInS-Net address three key challenges in joint medical image interpolation and segmentation. First, the spatially continuous position interpolation module solves the arbitrary-position modeling problem for interpolated frames, generating a physical-spacing-aware position prior and correcting a linear feature blend through a gated residual at every scale. Unlike fixed-frequency positional encoding used in existing methods, this module incorporates the physical spacing of anisotropic volumes and guarantees exact endpoint recovery by construction, significantly improving the accuracy of lesion boundary reconstruction in interpolated frames. Second, the prototype-based task mutual interaction module establishes lightweight bidirectional communication. Segmentation prototypes constrain interpolation to preserve structural integrity, while interpolation prototypes convey inter-slice transition cues that guide segmentation to perceive z-axis deformation. Because the two branches meet only through a few prototypes, the module introduces negligible computational overhead while delivering consistent performance gains. Third, the multi-scale task-cooperative decoder selectively routes cross-task spatial information through learned sigmoid gates at each upsampling scale, refining local details and boundaries without extra supervision. Ablation studies show that these components contribute consistently to both tasks, especially improving the boundary-sensitive recall metric.

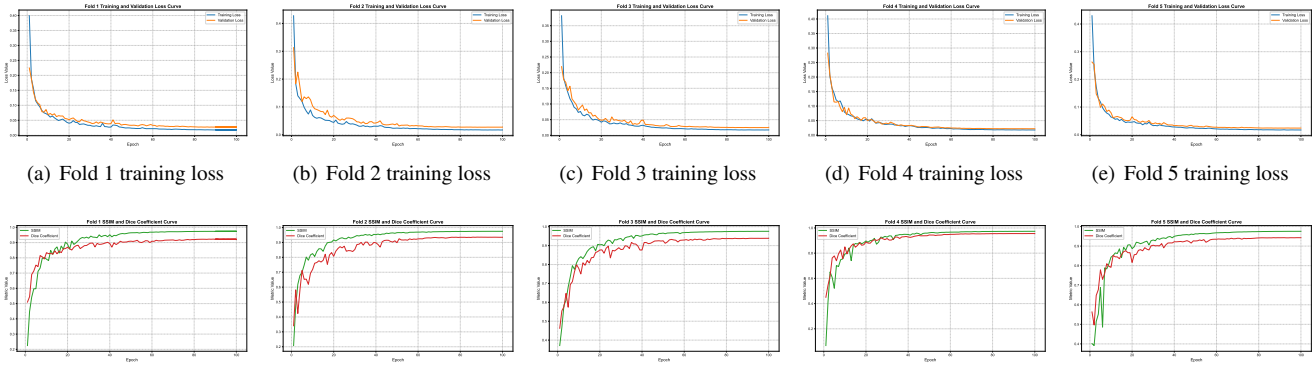
Figure 7 displays five-fold cross-validation training curves on the MSD Heart dataset. The multi-task total training loss steadily declines and converges while validation interpolation SSIM and segmentation Dice metrics keep rising and stabilize at high values throughout model training.

From a clinical standpoint, the joint optimization framework brings substantial practical value. The improved boundary precision of both interpolated images and segmentation masks benefits radiotherapy planning and longitudinal monitoring, where small contour deviations may affect target definition and dosimetry. The reduced inference latency and memory footprint also enable deployment on resource-limited clinical workstations, facilitating routine clinical workflow integration. The strong performance across diverse anatomical regions and modalities further suggests broad applicability for different clinical scenarios.

Nevertheless, this study has several limitations. First, all evaluations are conducted on publicly available benchmark datasets, which do not fully represent the diversity of real-world clinical data acquired under varying scanning protocols and vendor distributions. External validation on independent, institution-specific clinical datasets is required to further confirm generalizability. Second, the current framework operates on 2D slices, which does not fully exploit 3D volumetric context for spatial consistency. Future work will extend the architecture to pure 3D volumetric processing to further enhance inter-slice continuity. Third, the model performance on extremely small lesions remains inferior to that on medium and large lesions, consistent with known limitations of overlap-based metrics for small volumes. Future studies could incorporate size-aware sampling strategies or uncertainty estimation to improve performance on low-volume cases.

7. Conclusion

This paper presents CoInS-Net, an advanced multi-task framework for joint medical image interpolation and segmentation that integrates spatially continuous position interpolation, prototype-based bidirectional task interaction, and a multi-scale task-cooperative decoder. The results from extensive evaluation on four diverse medical imaging benchmarks showcase the model's superiority in terms of both reconstruction accuracy and segmentation precision, making it a significant advancement in multi-task medical image analysis. CoInS-Net consistently outperforms existing single-task methods across various anatomical regions and modalities, demonstrating its ability to exploit cross-task complementarity and providing clinically relevant image analysis results. The framework's robust performance across diverse datasets, coupled with its high computational efficiency, indicates potential for real-world clinical deployment. Future work will focus on extending the model to 3D volumetric processing and validating performance on multi-center clinical datasets.



(a) Fold 1 training loss (b) Fold 2 training loss (c) Fold 3 training loss (d) Fold 4 training loss (e) Fold 5 training loss

(f) Fold 1 validation metrics (g) Fold 2 validation metrics (h) Fold 3 validation metrics (i) Fold 4 validation metrics (j) Fold 5 validation metrics

Figure 7: Multi-task training convergence curves of five-fold cross-validation on the MSD Heart dataset. (a)–(e) Total training loss curves across the five folds; (f)–(j) Validation curves of interpolation SSIM and segmentation Dice across the five folds.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was not supported by any fundation.

References

[1] S. Lam et al. Current and future perspectives on ct screening for lung cancer: A road map for 2023-2027 from the iaslc. *Journal of Thoracic Oncology*, 19(1):36–51, January 2024.

[2] Zhicheng Zhang, Xiaokun Liang, Xu Dong, Yaoqin Xie, and Guohua Cao. A sparse-view ct reconstruction method based on combination of densenet and deconvolution. *IEEE Transactions on Medical Imaging*, 37(6):1407–1417, jun 2018.

[3] Zhikai Ruan, Canxu Song, Pengfei Xu, Chaoyu Wang, Jing Zhao, Meng Chen, Suoni Li, Qiang Su, Xiaozhen Zhuo, Yue Wu, Mingxi Wan, and Diya Wang. Multiparametric ultrasound breast tumors diagnosis within bi-rads category 4 via feature disentanglement and cross-fusion. *IEEE Transactions on Medical Imaging*, 44(7):3064–3075, jul 2025.

[4] Kwang-Hyun Uhm, Hyunjun Cho, Sung-Hoo Hong, and Seung-Won Jung. An anisotropic cross-view texture transfer with multi-reference non-local attention for ct slice interpolation. *IEEE Transactions on Medical Imaging*, 45(1):336–349, jan 2026.

[5] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, December 2020.

[6] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*, 9:82031–82057, 2021.

[7] Y. Cai et al. Swin unet3d: a three-dimensional medical image segmentation network combining vision transformer and convolution. *BMC Medical Informatics and Decision Making*, 23(1), February 2023.

[8] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 272–284. 2022.

[9] Zhaotao Wu, Jia Wei, Jiabing Wang, and Rui Li. Slice imputation: Multiple intermediate slices interpolation for anisotropic 3d medical image segmentation. *Computers in Biology and Medicine*, 147:105667, jun 2022.

[10] Wing Keung Cheung, Ashkan Pakzad, Nesrin Mogulkoc, et al. Interpolation-split: a data-centric deep learning approach with big interpolated data to boost airway segmentation performance. *Journal of Big Data*, 11(1), aug 2024.

[11] Muhammad Sarmad, Leonardo Carlos Ruspini, and Frank Lindseth. Sit-sr 3d: Self-supervised slice interpolation via transfer learning for 3d volume super-resolution. *Pattern Recognition Letters*, 166:97–104, feb 2023.

[12] Tzung-Chi Huang, Geoffrey Zhang, Thomas Guerrero, George Starkschall, Kan-Ping Lin, and Ken Forster. Semi-automated ct segmentation using optic flow and fourier interpolation techniques. *Computer Methods and Programs in Biomedicine*, 84(2–3):124–134, oct 2006.

[13] Y. Zhao, X. Wang, T. Che, G. Bao, and S. Li. Multi-task deep learning for medical image computing and analysis: A review. *Computers in Biology and Medicine*, 153:106496, February 2023.

[14] Z. Marinov, S. Reiß, D. Kersting, Jens Kleesiek, and Rainer Stiefelhagen. Mirror u-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2275–2285, October 2023.

[15] Z. Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

[16] Eduard Schreibmann, George T. Y. Chen, and Lei Xing. Image interpolation in 4d ct using a bspline deformable registration model. *International Journal of Radiation Oncology Biology Physics*, 64(5):1537–1550, apr 2006.

[17] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, volume 11045, pages 3–11. 2018.

[18] Hao Bai, Xibo Zhou, Yue Zhao, Yandong Zhao, and Qiaoling Han. Apflownet: An inter-layer interpolation approach for soil ct images based on cnn and bidirectional optical flow. *Soil and Tillage Research*, 238:106024, feb 2024.

[19] Lianying Chao, Zhiwei Wang, Haobo Zhang, Wei Xu, Peng Zhang, and Qiang Li. Sparse-view cone beam ct reconstruction using dual cnns in projection domain and image domain. *Neurocomputing*, 493:536–547, jul 2022.

[20] Pengxin Yu, Haoyue Zhang, Dawei Wang, et al. Spatial resolution enhancement using deep learning improves chest disease diagnosis based on thick slice ct. *npj Digital Medicine*, 7(1), nov 2024.

[21] N. Que, X. Wang, J. Chen, Y. Jiang, and C. Li. Adaptive spatial transcriptomics interpolation via cross-modal cross-slice modeling.

In *Lecture Notes in Computer Science*, pages 45–54, September 2025.

[22] Y. Guo, L. Bi, E. Ahn, D. Feng, Q. Wang, and J. Kim. A spatiotemporal volumetric interpolation network for 4d dynamic medical image. arXiv (Cornell University), June 2020.

[23] C. Peng, Wei-An Lin, H. Liao, R. Chellappa, and S. Kevin Zhou. Saint: Spatially aware interpolation network for medical slice synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7747–7756, 6 2020.

[24] Q. et al. Wu. An arbitrary scale super-resolution approach for 3d mr images via implicit neural representation. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1004–1015, February 2023.

[25] X. et al. Wang. Spatial attention-based implicit neural representation for arbitrary reduction of mri slice spacing. *Medical Image Analysis*, 94:103158, March 2024.

[26] H. Song, X. Mao, J. Yu, Q. Li, and Y. Wang. I3net: Inter-intra-slice interpolation network for medical slice synthesis. *IEEE Transactions on Medical Imaging*, 43(9):3306–3318, Apr. 2024.

[27] X. Yan, H. Tang, S. Sun, H. Ma, D. Kong, and X. Xie. After-unet: Axial fusion transformer unet for medical image segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3270–3280, January 2022.

[28] Xueren Zhang, Xianghong Wang, and Tianye Niu. Ct image segmentation using frequency domain feature-assisted selective long memory state space model. *Sensing and Imaging*, 26(1), may 2025.

[29] X. Zhang, X. Wang, and T. Niu. Ct image segmentation using frequency domain feature-assisted selective long memory state space model. *Sensing and Imaging*, 26(1), May 2025.

[30] J. Yang, P. Qiu, Y. Zhang, D. S. Marcus, and A. Sotiras. D-net: Dynamic large kernel with dynamic feature fusion for volumetric medical image segmentation. *Biomedical Signal Processing and Control*, 113:108837, March 2026.

[31] Kaiqi Dong, Peijun Hu, Yan Zhu, et al. Attention-enhanced multiscale feature fusion network for pancreas and tumor segmentation. *Medical Physics*, 51(12):8999–9016, sep 2024.

[32] K. Dong et al. Attention-enhanced multiscale feature fusion network for pancreas and tumor segmentation. *Medical Physics*, 51(12):8999–9016, September 2024.

[33] N. Shen et al. Multi-organ segmentation network for abdominal ct images based on spatial attention and deformable convolution. *Expert Systems with Applications*, 211:118625, January 2023.

[34] M. et al. Wang. Ba-unet: A boundary augmented segmentation network for cervical cancer radiotherapy. *Journal of Imaging Informatics in Medicine*, April 2026.

[35] X. Liu, Y. Hu, J. Chen, and K. Li. Shape and boundary-aware multi-branch model for semi-supervised medical image segmentation. *Computers in Biology and Medicine*, 143:105252, April 2022.

[36] Z. Wu, J. Wei, J. Wang, and R. Li. Slice imputation: Multiple intermediate slices interpolation for anisotropic 3d medical image segmentation. *Computers in Biology and Medicine*, 147:105667, August 2022.

[37] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[38] Sebastian Ruder. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098, 2017.

[39] G. Dai et al. Multi-task learning network for medical image analysis guided by lesion regions and spatial relationships of tissues. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, January 2025.

[40] Pedro Esteban Chavarrias Solano, Andrew Bulpitt, Venkataraman Subramanian, and Sharib Ali. Multi-task learning with cross-task consistency for improved depth estimation in colonoscopy. *Medical Image Analysis*, 99:103379, jan 2025.

[41] Y. Sun, Y. Yang, and N. Bao. Tasc-swinmt: Task-adaptive synergistic cross-task swin multi-task framework for ct and mri image interpolation and segmentation. *Tomography*, 12(6):80, May 2026.

[42] P. E. Chavarrias Solano, A. Bulpitt, V. Subramanian, and S. Ali. Multi-task learning with cross-task consistency for improved depth estimation in colonoscopy. *Medical Image Analysis*, 99:103379, January 2025.

[43] G. Marullo, L. Tanzi, L. Ulrich, F. Porpiglia, and E. Vezzetti. A multi-task convolutional neural network for semantic segmentation and event detection in laparoscopic surgery. *Journal of Personalized Medicine*, 13(3):413–413, feb 2023.

[44] M. Nohel, C. Ulrich, J. Suprijadi, T. Wald, and K. Maier-Hein. Unified framework for foreground and anonymization area segmentation in ct and mri data. In *Informatik aktuell*, pages 242–247. 2025.

[45] H. Chen et al. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE Transactions on Medical Imaging*, 36(12):2524–2535, December 2017.

[46] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003, 2016.

[47] J. Ma, Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939, 7 2018.

[48] H. Tang, J. Liu, M. Zhao, and X. Gong. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems*, pages 22–31, 9 2020.

[49] G. Li et al. Rethinking multi-contrast MRI super-resolution: Rectangle-window cross-attention transformer and arbitrary-scale upsampling. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21173–21183, October 2023.

[50] C. Peng, W.-A. Lin, H. Liao, R. Chellappa, and S. K. Zhou. SAINT: Spatially aware interpolation network for medical slice synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7747–7756, June 2020.

[51] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. In *Computer Vision – ECCV 2018*, pages 294–310, 2018.

[52] K.-H. Uhm, H. Cho, S.-H. Hong, and S.-W. Jung. An anisotropic cross-view texture transfer with multi-reference non-local attention for ct slice interpolation. *IEEE transactions on medical imaging*, 45(1):336–349, Jan. 2026.

[53] W. Fang et al. CycleINR: Cycle implicit neural representation for arbitrary-scale volumetric super-resolution of medical data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, December 2024.

[54] W. Li et al. Sfcli-net: Spatial-frequency collaborative learning interpolation network for computed tomography slice synthesis. *Expert Systems with Applications*, 272:126602, Jan. 2025.

[55] Z. Xing et al. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2025.

[56] T. Chen, C. Wang, Z. Chen, Y. Lei, and H. Shan. Hidiff: Hybrid diffusion framework for medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(10):3570–3583, Oct. 2024.

[57] P. Wang et al. Accurate airway tree segmentation in ct scans via anatomy-aware multi-class segmentation and topology-guided iterative learning. *IEEE Transactions on Medical Imaging*, pages 1–1, Jan. 2024.

[58] J. Wu, X. Liu, G. Wang, and S. Zhang. Sictta: Single image continual test time adaptation for medical image segmentation. *Medical Image Analysis*, 108:103859, Feb. 2026.

[59] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Learning Representations (ICLR)*, Feb 2018. Accessed May 15, 2026.

[60] S. Liu, E. J. Johns, and A. J. Davison. End-to-end multi-task learning with attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.