Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.087647

ARTICLE

Scale Ladder Consistency for Structure-Aware Multimodal Representation Learning in 3D Medical Image Segmentation

Weiqing Liu[1,#], **Bin Li**[1,*,#], **Lianfang Tian**[1] and **Qianhui Qiu**[2]

[1]School of Automation Science and Engineering, South China University of Technology, Guangzhou, China
[2]Department of Otolaryngology, Head and Neck Surgery, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, China

*Corresponding Author: Bin Li. Email: binlee@scut.edu.cn

[#]These authors contributed equally to this work

Received: 20 June 2026; Accepted: 12 August 2026; Published: 28 August 2026

ABSTRACT: Self-supervised representation learning can reduce the dependence of three-dimensional (3D) medical image segmentation on dense voxel annotations. In multimodal 3D medical imaging, intensity-reconstruction pre-training provides dense appearance supervision but does not explicitly distinguish the structural regions that determine segmentation boundaries and small targets. A second mismatch arises in scale learning: encoder-decoder networks provide multi-scale feature maps, but they do not explicitly supervise how fine anatomical structures weaken or persist across neighboring scales. To address these mismatches, this study proposes Scale Ladder Consistency (SLC), a structure-aware self-supervised representation learning framework for multimodal 3D medical image segmentation. SLC combines Scale-Space Structural Reconstruction (SSR), Hybrid Mask, and Scale Ladder (SL) in its structural pre-training path. SSR replaces intensity recovery with structure prediction, Hybrid Mask increases supervision on fine structural regions, and SL learns neighboring-scale structural transitions through bidirectional prediction. Subset-to-Full Regularization (S2F) further stabilizes case-level representations during pre-training. Experimental results on the Brain Tumor Segmentation 2019 (BraTS19) and carotid artery datasets demonstrate that SLC consistently outperforms matched scratch fine-tuning and achieves competitive performance against recent segmentation and self-supervised methods. These results indicate that structure-aware and cross-scale self-supervised objectives can provide effective representations for multimodal 3D medical segmentation.

KEYWORDS: Representation learning; self-supervised learning; multimodal imaging; 3D medical image segmentation; scale-space structural reconstruction; medical image analysis

1 Introduction

Accurate 3D medical image segmentation supports image-guided diagnosis, treatment planning, and quantitative disease assessment. Self-configuring convolutional pipelines, Transformer-based volumetric models, and modernized convolutional architectures have achieved strong performance when task-specific annotations are sufficient [1–3]. Dense voxel labels, however, are costly and require clinical expertise. The burden is particularly high for thin vessels, small lesions, and complex boundaries. Self-supervised learning (SSL) offers a practical way to use unlabeled multimodal volumes before downstream fine-tuning.

Masked image modeling (MIM) is a common SSL paradigm for 3D medical images [4–6]. It hides image regions and trains a network to recover the missing content. Recent studies have expanded this paradigm through larger pre-training resources, structure-aware objectives, and complementary corruption

processes [7–10]. These advances improve transferability, but they also sharpen a task-specific question: what should be predicted to emphasize the sparse anatomical structures required by the downstream task?

The first mismatch concerns the prediction target. Intensity reconstruction rewards appearance recovery across the full volume, whereas segmentation decisions are concentrated around boundaries, vessels, and lesion interfaces. These structures occupy a small spatial fraction, so their relevance is not explicitly represented by an intensity target. Topology- and connectivity-aware segmentation studies likewise show that thin structures and coherent boundaries require structural cues beyond region-level overlap [11,12].

The same mismatch affects the spatial allocation of masked supervision. Purely random masking may expose few voxels from a sparse vessel or a small lesion boundary. Medical MIM methods have therefore explored masked multi-view learning, hybrid masking, and lesion-focused sampling [13–16]. Their findings indicate that both the target content and the sampled locations influence the learned representation.

The second mismatch concerns scale learning. A structure-guided mask can determine where supervision is applied, but it does not describe how a structure changes across resolution levels. Hierarchical segmentation and pre-training methods benefit from multi-resolution features, cross-level alignment, or feature-pyramid decoding [17–19]. Nevertheless, the presence or fusion of multi-scale feature maps does not directly supervise their structural transitions. A boundary visible at a fine scale may weaken after smoothing or downsampling. A scale-aware pre-training objective should therefore learn what persists and what disappears between neighboring scales.

Fig. 1 summarizes these two scientific problems. Intensity reconstruction provides dense appearance supervision without explicitly identifying segmentation-relevant regions. The scale-feature examples show that structural responses change across decoder scales, motivating explicit neighboring-scale structural prediction.

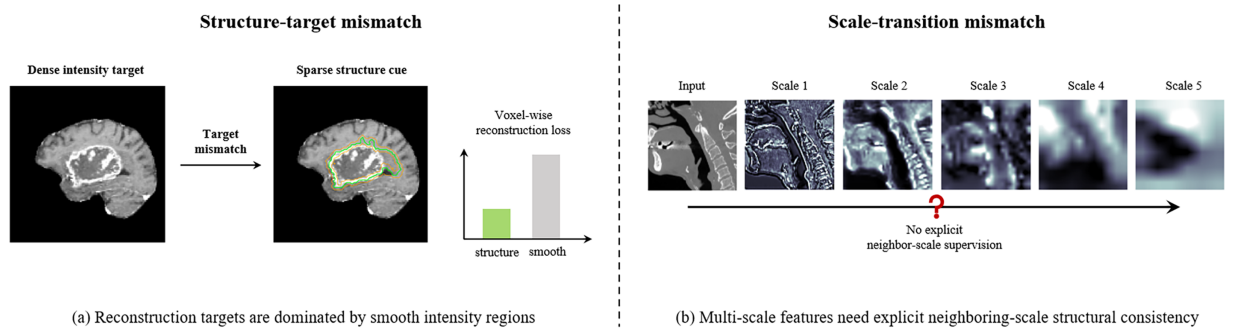

Figure 1: Motivation of SLC. (**a**) Voxel-wise intensity reconstruction distributes loss over dense image regions, while segmentation-relevant boundaries and small structures occupy a sparse fraction. (**b**) Decoder features become progressively smoother from Scale 1 to Scale 5; the arrow and question mark indicate that their neighboring-scale structural transitions lack explicit supervision. SLC addresses these two problems with structure-aware reconstruction targets and bidirectional adjacent-scale prediction.

To address these two mismatches, this study proposes Scale Ladder Consistency (SLC), a structure-aware self-supervised representation learning framework for multimodal 3D medical image segmentation. SLC combines Scale-Space Structural Reconstruction (SSR), Hybrid Mask, and Scale Ladder (SL) in its structural pre-training path. SSR replaces intensity recovery with three image-derived structural targets. Hybrid Mask allocates part of the mask budget according to the fine-detail response while retaining random blocks

for contextual learning. SL learns transitions between adjacent decoder scales through bidirectional fine-to-coarse and coarse-to-fine prediction. Subset-to-Full Regularization (S2F) uses an exponential moving average (EMA) teacher as an auxiliary representation stabilizer.

The framework uses a modality-set encoder-decoder. Modality-specific stems process the inputs, a shared encoder extracts spatial features, and a token-set Transformer represents the visible modalities. Token-conditioned decoding then injects multimodal context into the spatial feature maps. This formulation supports variable modality sets without requiring strict registration between heterogeneous modality spaces. The main contributions are summarized as follows:

(1) SLC is developed as a structure-aware multimodal self-supervised representation learning framework for 3D medical image segmentation.
(2) SSR and Hybrid Mask redirect pre-training supervision toward sparse boundaries and fine anatomical structures through complementary structural targets and fine-detail-guided mask allocation.
(3) SL learns neighboring-scale structural transitions through bidirectional prediction, while S2F serves as an auxiliary stabilizer during pre-training.

2 Related Work

2.1 Self-Supervised Representation Learning for 3D Medical Image Segmentation

Early 3D medical SSL used restoration, semantic self-discovery, and transformation prediction to learn transferable features from unlabeled volumes [20,21]. Contrastive learning later improved label efficiency by comparing global and local anatomical features [22]. Other volumetric methods use geometric similarity, volume-level contrast, or mixed two-dimensional and 3D data to broaden the pre-training signal [23–25].

Restoration-based SSL commonly uses masked reconstruction. Masked autoencoders established masked reconstruction as an effective strategy for visual representation learning in natural images [4]. Medical extensions include Swin Transformer pre-training, volumetric MIM, masked multi-view learning, and cross-dimensional pseudo-3D transformation [5,6,13,26]. Haghighi et al. [27] further combined discriminative, restorative, and adversarial objectives to capture complementary medical-image information.

Recent work has expanded both the scale and the design of 3D medical SSL. OpenMind provides a large public brain magnetic resonance imaging resource and a standardized benchmark for several pre-training methods [7]. Structure-aware Semantic Discrepancy and Consistency (S2DC) aligns patch-level semantic discrepancy with neighborhood-based structural consistency [8]. MAsk-guided Self-Supervision (MASS) uses automatically generated class-agnostic masks as structural pre-training targets [9]. Masked-Diffusion Autoencoders (MDAE) combine masked reconstruction with visible-region diffusion denoising to learn structural and textural information [10]. MedCSS introduces causal regularization and hierarchical feature consistency for 3D representation learning [28].

These studies establish the value of richer data and complementary objectives. The prediction target remains important for segmentation-oriented pre-training. The proposed SLC framework uses image-derived scale-space responses for structural reconstruction and SL to supervise transitions between adjacent decoder scales.

2.2 Structure-Aware Visual and Spatial Representation Learning

Medical segmentation is sensitive to boundaries, connectivity, and small foreground regions. The centerline Dice (clDice) loss preserves topology in tubular structures [11], while directional connectivity modeling improves anatomical continuity [12]. Topological interaction learning represents containment and exclusion constraints in multi-class segmentation [29]. For small lesions, the Clinical Diagnosis-Inspired Non-Salient

Small Tumor Segmentation framework (CDI-NSTSEG) follows a screening-to-refinement process [30]. Collaborative learning of dynamic and static information further combines within-slice structure with inter-slice variation [31].

Structure-aware learning addresses the same concern in supervised and self-supervised settings. BoundaryMIM incorporates boundary-oriented masked reconstruction into supervised segmentation training [32]. Hybrid masked image modeling combines pixel-, region-, and sample-level supervision for 3D medical pre-training [14]. Adaptive and lesion-focused masking strategies place more pre-training signal around boundaries or lesion-rich regions [15,16,33].

S2DC and MASS are particularly relevant recent approaches. S2DC learns semantic discrepancy between patches and consistency within inferred structures [8]. MASS treats automatically generated masks as class-agnostic structural targets [9]. In contrast, SLC derives three continuous responses from the image scale space and predicts how structural responses change between neighboring decoder levels. The fine-detail response also guides part of the masking process, while random regions preserve contextual coverage.

2.3 Multi-Scale Representation Learning in Medical Image Segmentation

Multi-scale representation learning is fundamental to medical image segmentation. Self-configuring U-shaped pipelines remain strong biomedical baselines [1]. Volumetric Transformers capture long-range dependencies through patch-based sequence modeling [2], while modernized convolutional blocks provide an effective alternative for data-scarce tasks [3]. Hierarchical models further combine local and global attention or align representations across resolution levels [17,18,34].

Recent segmentation architectures refine this feature-fusion perspective. M^3-TransUNet combines spatial priors with multi-scale gating [35]. Multi-Scale Selective Downsampling and Non-Adjacent Layers Guidance (SDNAL-Seg) uses adaptive downsampling and guidance between non-adjacent layers [36]. The Context-Aware Adaptive Progressive Network (CA^2PNet) combines context-aware attention with progressive dilated convolutions [37]. SegDINO reorganizes intermediate self-supervised tokens into a pyramid and applies scale-aware decoding [19]. These methods improve multi-scale extraction or fusion within a segmentation architecture.

Scale-aware pre-training provides a complementary view. Swin Transformer pre-training, masked multi-view learning, and cross-level alignment exploit hierarchical representations before fine-tuning [5,13,18]. MedCSS also aligns intermediate and high-level feature distributions [28]. However, the existence, alignment, or fusion of multi-scale features does not specify how an anatomical response should weaken or persist from one level to the next. SL therefore uses bidirectional adjacent-scale prediction as an explicit structural pre-training signal.

2.4 Multimodal Learning with Incomplete Observations

Multimodal medical image analysis aims to exploit complementary anatomical and tissue information across imaging modalities, but incomplete inputs, modality heterogeneity, and registration uncertainty remain common. Many incomplete-modality methods assume co-registered inputs. The multimodal medical Transformer (mmFormer) models intramodal and intermodal dependencies for incomplete brain tumor segmentation [38]. Multimodal representation learning with missing modalities (M3AE) combines modality dropout with patch masking [39]. The Missing-as-Masking cross-modal feature reconstruction method (M3FeCon) reconstructs absent modalities in feature space [40].

Teacher-student and distillation-style training provide another way to stabilize representations when the input view is incomplete. Momentum or EMA teachers provide slowly updated targets in SSL [41,42].

In multimodal medical segmentation, cross-modality collaboration also supports semi-supervised learning under scarce labels and modality misalignment [43]. These works motivate the use of a more complete view as a stable reference. In the proposed framework, S2F implements this teacher-student principle as an auxiliary regularizer alongside scale-space structural pre-training.

This study uses a modality-set formulation for variable multimodal inputs. Each available image contributes one element to the case-level set, and aggregation is performed over the elements present in that case. S2F uses the complete case as a stable reference. It aligns the case-level representation of the subset-view student with that of the full-view EMA teacher. SSR predicts voxel-level structural responses. SL supervises transitions between adjacent decoder scales. Hybrid Mask selects the locations used for reconstruction supervision.

3 Method

3.1 Overview and Problem-Driven Design

SLC addresses two mismatches between reconstruction-based pre-training and segmentation-oriented representation learning. First, intensity recovery does not explicitly identify segmentation-relevant structures. SSR instead defines structure-oriented targets, and Hybrid Mask uses the fine-detail response to sample more anatomical detail. Second, feature pyramids do not explicitly supervise structural transitions across scales. SL introduces bidirectional prediction between adjacent decoder scales. S2F serves as an auxiliary case-level stabilizer.

Let a training case be a variable modality set $\mathcal{C} = \{(\mathbf{X}_j, m_j)\}_{j=1}^{n}$, where n is the number of available images, $\mathbf{X}_j \in \mathbb{R}^{1\times D\times H\times W}$ is a 3D image, and m_j is its modality identifier. During pre-training, one image is selected as the anchor $\mathbf{X}_a$. The student view consists of the masked anchor and a modality-dropped subset of the remaining same-case images, denoted by $\mathcal{C}_{\text{ctx}}$. The unmasked anchor is excluded from the student input. A binary hybrid mask $\mathbf{M}_a \in \{0,1\}^{1\times D\times H\times W}$ is generated as described in Section 3.4. The masked anchor is $\widetilde{\mathbf{X}}_a = \mathbf{X}_a \odot (1 - \mathbf{M}_a)$, where $\odot$ denotes element-wise multiplication. The student network receives $\widetilde{\mathbf{X}}_a$ and the retained context $\mathcal{C}_{\text{ctx}}$. It returns a case-level representation $\mathbf{g}_{\text{sub}}$, a three-channel structure prediction $\widehat{\mathbf{S}}_a$, and token-conditioned scale features at $L = 5$ scales:

$$\mathbf{g}_{\text{sub}}, \widehat{\mathbf{S}}_a, \{\mathbf{F}^{(i)}\}_{i=1}^{L} = G_\theta(\widetilde{\mathbf{X}}_a, \mathcal{C}_{\text{ctx}}). \tag{1}$$

Here θ denotes the learnable parameters of the student network. Throughout this section, the scale index is written as a superscript: $i = 1$ denotes the finest decoder scale and $i = L$ denotes the coarsest scale. Architecture width and crop settings are reported in Section 4.2.

The pre-training objective combines S2F loss, SSR loss, and SL loss:

$$\mathcal{L}_{\text{pre}} = \lambda_{\text{s2f}}\mathcal{L}_{\text{s2f}} + \lambda_{\text{ssr}}\mathcal{L}_{\text{ssr}} + \lambda_{\text{sl}}\mathcal{L}_{\text{sl}}, \tag{2}$$

with $\lambda_{\text{s2f}} = 0.3$, $\lambda_{\text{ssr}} = 0.5$, and $\lambda_{\text{sl}} = 0.5$. The SSR loss reconstructs voxel-level structural responses, the SL loss supervises structural transitions between neighboring scales, and the S2F loss stabilizes case-level representations using the full-view EMA teacher. The overall workflow is shown in Fig. 2.

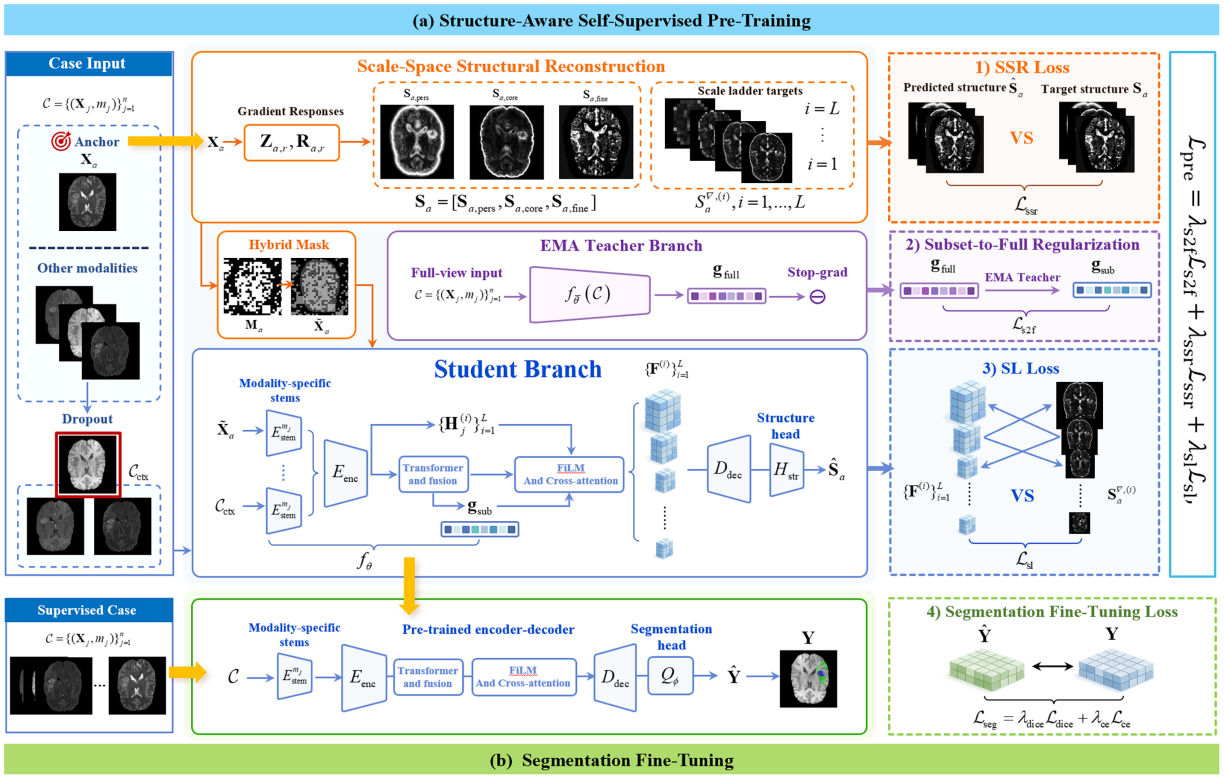

Figure 2: Overview of SLC. (**a**) Self-supervised pre-training uses a masked student view and a full-view EMA teacher view from a variable modality set. SSR supervises voxel-level structural reconstruction, SL links neighboring structural scales, and S2F stabilizes the case-level representation. (**b**) Supervised fine-tuning attaches a segmentation head to the pre-trained encoder-decoder.

3.2 Modality-Set Encoder-Decoder

The backbone must accept a variable set of imaging modalities while producing an anchor-centered decoder representation for structure prediction. Each image in $\mathcal{C}$ is first encoded independently by a modality-specific stem, $\mathbf{H}_j^{(1)} = E_{\text{stem}}^{m_j}(\mathbf{X}_j)$, where $\mathbf{H}_j^{(1)}$ is the finest feature of image j. Shared encoder blocks then produce a five-level feature pyramid $\{\mathbf{H}_j^{(i)}\}_{i=1}^{L}$:

$$\mathbf{H}_j^{(i)} = E_{\text{enc}}^{(i)}(\mathbf{H}_j^{(i-1)}), \quad i = 2, \ldots, L, \tag{3}$$

where $E_{\text{enc}}^{(i)}$ denotes the shared encoder stage at scale i. When $j = a$, these features form the anchor pyramid $\{\mathbf{H}_a^{(i)}\}_{i=1}^{L}$ used by the reconstruction decoder. This shared hierarchy keeps modality-specific processing at the input stem while mapping all visible images into a common multi-scale feature space. The bottleneck feature is average-pooled and combined with a modality embedding to form $\mathbf{t}_j = \text{GAP}(\mathbf{H}_j^{(L)}) + \mathbf{e}_{m_j}^{\text{mod}}$. Here, GAP denotes global average pooling, and $\mathbf{e}_{m_j}^{\text{mod}}$ encodes the identity of modality m_j. Let $\mathcal{V}$ denote the visible input indices in the case. The visible modality tokens are processed by a Transformer encoder, $\mathbf{T} = \{\mathbf{T}_j\}_{j\in\mathcal{V}} = \text{Transformer}(\{\mathbf{t}_j\}_{j\in\mathcal{V}})$, with padding masks applied to absent modalities. Missing-aware fusion obtains a case-level representation by learned token aggregation:

$$\mathbf{g} = \mathrm{LN}\left(\sum_{j\in\mathcal{V}} \alpha_j \mathbf{T}_j\right), \quad \alpha_j = \frac{\exp(q(\mathbf{T}_j))}{\sum_{k\in\mathcal{V}} \exp(q(\mathbf{T}_k))}, \tag{4}$$

where $q(\cdot)$ is a learned scalar scoring function and LN denotes layer normalization. When this representation is computed from the student's masked and modality-dropped view, it is denoted by $\mathbf{g}_{\mathrm{sub}}$; when it is computed from the teacher's full view, it is denoted by $\mathbf{g}_{\mathrm{full}}$.

For SLC reconstruction, the decoder uses the anchor pyramid $\{\mathbf{H}_a^{(i)}\}_{i=1}^{L}$. The anchor mask is resized to each spatial scale and applied to the first four levels:

$$\widetilde{\mathbf{H}}_a^{(i)} = \mathbf{H}_a^{(i)} \odot \left(1 - \mathrm{Interp}_{s^{(i)}}(\mathbf{M}_a)\right), \quad i = 1, \ldots, L-1, \tag{5}$$

and $\widetilde{\mathbf{H}}_a^{(L)} = \mathbf{H}_a^{(L)}$. Here $s^{(i)}$ denotes the spatial size at scale i, shared by $\mathbf{H}_j^{(i)}$ and $\mathbf{F}^{(i)}$, and the mask is resized by nearest-neighbor interpolation. The global representation $\mathbf{g}$ modulates the scale features through Feature-wise Linear Modulation (FiLM). Token-to-feature cross-attention then injects the modality-set context from $\mathbf{T}$ into all but the finest scale:

$$\begin{cases} \mathbf{F}^{(1)} = \Gamma^{(1)}(\widetilde{\mathbf{H}}_a^{(1)}, \mathbf{g}), \\ \mathbf{F}^{(i)} = \mathcal{A}^{(i)}\left(\Gamma^{(i)}(\widetilde{\mathbf{H}}_a^{(i)}, \mathbf{g}), \mathbf{T}\right), \quad i = 2, \ldots, L, \end{cases} \tag{6}$$

where $\Gamma^{(i)}(\cdot)$ denotes scale-specific FiLM modulation and $\mathcal{A}^{(i)}(\cdot)$ denotes token-to-feature cross-attention. The features $\{\mathbf{F}^{(i)}\}_{i=1}^{L}$ are used by SL in Eq. (14). They are also passed to the decoder and structure head:

$$\widehat{\mathbf{S}}_a = H_{\mathrm{str}}\left(D_{\mathrm{dec}}(\mathbf{F}^{(1)}, \ldots, \mathbf{F}^{(L)})\right). \tag{7}$$

Here D_{dec} denotes the reconstruction decoder, and H_{str} denotes the structure prediction head. This prediction is supervised by the SSR loss in Eq. (10).

3.3 Scale-Space Structural Reconstruction Loss (SSR Loss)

Scale-Space Structural Reconstruction (SSR) changes the target from intensity recovery to explicit structure prediction. It constructs three structural targets from the unmasked anchor image and supervises the masked student prediction. Given $\mathbf{X}_a$, SSR computes gradient responses after smoothing at several radii. For each $r \in \mathcal{R} = \{0, 1, 2, 4\}$,

$$Z_{a,r} = \mathrm{zscore}\left(\|\nabla(K_r * X_a)\|_2\right), \qquad R_{a,r} = \sigma\left(\frac{Z_{a,r} - 0.75}{0.50}\right), \tag{8}$$

where K_r denotes 3D average-pooling smoothing with kernel size $2r+1$, and K_0 is the identity operator. $\sigma(\cdot)$ is the sigmoid function, and the z-score is computed per volume and clipped to $[-6, 6]$ before the sigmoid.

The SSR target contains three channels:

$$\begin{cases} \mathbf{S}_{a,\mathrm{pers}} = \dfrac{1}{|\mathcal{R}|}\displaystyle\sum_{r\in\mathcal{R}} \mathbf{R}_{a,r}, \\ \mathbf{S}_{a,\mathrm{core}} = \min\limits_{r\in\mathcal{R}} \mathbf{R}_{a,r}, \\ \mathbf{S}_{a,\mathrm{fine}} = \max(\mathbf{R}_{a,0} - \mathbf{R}_{a,4}, 0). \end{cases} \tag{9}$$

The three channels are referred to as the persistence response, stable core, and fine detail. Persistence averages the responses across smoothing radii and highlights structures that remain visible over a range of spatial scales. Stable core retains the minimum response across radii and emphasizes locations with consistently strong responses at every scale. Fine detail isolates the response lost between the unsmoothed and strongly smoothed images, highlighting thin vessels, sharp interfaces, and other high-frequency local structures. The final SSR target is $\mathbf{S}_a = [\mathbf{S}_{a,\text{pers}}, \mathbf{S}_{a,\text{core}}, \mathbf{S}_{a,\text{fine}}]$.

Let $\mathbf{M}_a^{\mathbf{S}}$ denote $\mathbf{M}_a$ broadcast to the three structure channels. The SSR loss is computed only on masked voxels:

$$\mathcal{L}_{\text{ssr}} = \frac{\|\mathbf{M}_a^{\mathbf{S}} \odot (\widehat{\mathbf{S}}_a - \text{sg}(\mathbf{S}_a))\|_1}{\max(\|\mathbf{M}_a^{\mathbf{S}}\|_1, 1)}. \tag{10}$$

Here $\text{sg}(\cdot)$ denotes stop-gradient. The target is detached from the computation graph and serves as fixed structural supervision for the student decoder.

3.4 Hybrid Mask

SSR provides a fine detail response that highlights structures visible before strong smoothing. This response is useful for mask allocation because a purely random mask can spend much of its budget on homogeneous regions and provide limited supervision for sparse anatomical detail. Hybrid Mask uses the fine detail channel $\mathbf{S}_{a,\text{fine}}$ as a local prior for boundaries, small vessels, and other structures that are visible at fine scale but weaken after smoothing. At the same time, random block masking is retained so that representation learning still includes contextual recovery beyond high-response structural regions.

Let $\eta = 0.35$ be the overall mask ratio and $\rho = 0.70$ be the guided fraction. A voxel-level masking prior is first obtained from the fine detail response as $\mathbf{P}_a = \text{clip}(\mathbf{S}_{a,\text{fine}}/\tau + \delta, 0, 1)$, where $\tau = 2.0$ is the temperature and $\delta = 10^{-3}$ is a small uniform floor. The prior is then averaged over non-overlapping guided cells with side length p_g, yielding cell probabilities

$$\pi_c = \frac{\text{AvgPool}_{p_g}(\mathbf{P}_a)_c}{\sum_{c'} \text{AvgPool}_{p_g}(\mathbf{P}_a)_{c'} + \varepsilon}. \tag{11}$$

In Eq. (11), c indexes guided cells, AvgPool_{p_g} denotes 3D average pooling with both kernel size and stride set to p_g, and $\varepsilon = 10^{-8}$ is used for probability normalization. Guided cells are drawn without replacement according to π_c. With $N = \eta DHW$ target masked voxels, the guided branch selects approximately $\rho N / p_g^3$ cells and forms the binary mask $\mathbf{M}_a^{\text{guided}}$ after nearest-neighbor expansion to the image size. The remaining budget is filled by randomly placed block cells with side length p_r, producing $\mathbf{M}_a^{\text{block}}$. Overlaps between the two branches are merged by clipping:

$$\mathbf{M}_a = \text{clip}(\mathbf{M}_a^{\text{guided}} + \mathbf{M}_a^{\text{block}}, 0, 1). \tag{12}$$

Here $p_r = 16$ and $p_g = \max(4, p_r/2)$, giving guided cells of size 8^3 voxels and random block cells of size 16^3 voxels. This design concentrates structural reconstruction around sparse anatomical detail, while the random branch preserves contextual coverage beyond high-response regions.

3.5 Scale-Ladder Loss (SL Loss)

The decoder feature pyramid provides representations at multiple spatial scales, but these feature maps do not necessarily encode how anatomical structure changes from one scale to the next. Fine structures may

weaken or disappear after smoothing and downsampling, while coarse features may still provide context for inferring the corresponding fine-scale response. SL turns this transition into an explicit scale-aware representation learning task by predicting neighboring-scale structural targets in both directions. For each adjacent scale pair $(i, i+1)$, $i = 1, \ldots, L-1$, SL uses two lightweight predictors. The fine-to-coarse predictor $U^{(i)}$ maps information from $\mathbf{F}^{(i)}$ to the structural target at scale $i+1$. The coarse-to-fine predictor $D^{(i)}$ maps information from $\mathbf{F}^{(i+1)}$ back to the structural target at scale i.

The single-channel gradient target at spatial size $s^{(i)}$ is

$$\mathbf{S}_a^{\nabla,(i)} = \mathrm{Interp}_{s^{(i)}}\big(\mathrm{zscore}(\|\nabla \mathbf{X}_a\|_2)\big), \tag{13}$$

where $\mathrm{Interp}_{s^{(i)}}(\cdot)$ denotes interpolation to spatial size $s^{(i)}$; continuous feature and target maps are resized by trilinear interpolation. The ladder target is a single-channel normalized gradient-magnitude map at each decoder scale; SSR uses the three-channel target $\mathbf{S}_a$.

The bidirectional SL loss is

$$\mathcal{L}_{\mathrm{sl}} = \frac{1}{2(L-1)} \sum_{i=1}^{L-1} \Big[\ell_1\big(U^{(i)}\big(\mathrm{Interp}_{s^{(i+1)}}(\mathbf{F}^{(i)})\big), \mathbf{S}_a^{\nabla,(i+1)}\big) + \ell_1\big(D^{(i)}\big(\mathrm{Interp}_{s^{(i)}}(\mathbf{F}^{(i+1)})\big), \mathbf{S}_a^{\nabla,(i)}\big) \Big]. \tag{14}$$

Here $\ell_1(\cdot,\cdot)$ denotes the voxel-averaged L1 error over the target map. Each predictor contains three convolutional layers. A $1\times1\times1$ projection and a $3\times3\times3$ convolution are each followed by instance normalization and Gaussian error linear unit activation. A final $1\times1\times1$ layer maps the 32-channel hidden representation to one structural-response channel. With $L = 5$, the four adjacent scale pairs are modeled by eight direction-specific blocks. The loss encourages neighboring scales to predict each other's structural responses, making scale transition an explicit pre-training signal.

3.6 Subset-to-Full (S2F) Regularization with EMA Teacher

SSR, Hybrid Mask, and SL form the structural pre-training path. The SSR loss and SL loss act on the masked anchor and its token-conditioned scale features. Under modality dropout and anchor masking, a stable full-view reference supports the case-level representation. S2F aligns the subset-view student representation with a full-view teacher representation and serves as an auxiliary training stabilizer. Let $f_\theta(\cdot)$ denote the case-level representation branch of the student. For the subset view, $\mathbf{g}_{\mathrm{sub}} = f_\theta(\widetilde{\mathbf{X}}_a, \mathcal{C}_{\mathrm{ctx}})$. An EMA teacher with parameters $\bar{\theta}$ receives the full, unmasked modality set and provides the target representation $\mathbf{g}_{\mathrm{full}} = f_{\bar{\theta}}(\mathcal{C})$. The S2F loss aligns normalized student and teacher representations over a mini-batch of B cases, with b indexing a case in the mini-batch:

$$\mathcal{L}_{\mathrm{s2f}} = \frac{1}{B} \sum_{b=1}^{B} \left\| \frac{\mathbf{g}_{\mathrm{sub}}^{(b)}}{\|\mathbf{g}_{\mathrm{sub}}^{(b)}\|_2} - \mathrm{sg}\left(\frac{\mathbf{g}_{\mathrm{full}}^{(b)}}{\|\mathbf{g}_{\mathrm{full}}^{(b)}\|_2} \right) \right\|_2^2 . \tag{15}$$

At training step t, back-propagation first produces the updated student parameters θ_{t+1}. The teacher parameters are then updated by exponential moving average, $\bar{\theta}_{t+1} \leftarrow \mu\bar{\theta}_t + (1-\mu)\theta_{t+1}$, with $\mu = 0.996$. The teacher branch provides stop-gradient targets for the S2F loss.

3.7 Training and Fine-Tuning Protocol

During pre-training, the student receives one masked anchor modality and the retained same-case context $\mathcal{C}_{\mathrm{ctx}}$, whereas the EMA teacher receives the full, unmasked case $\mathcal{C}$. The student forward path in Eq. (1) produces $\mathbf{g}_{\mathrm{sub}}$, $\widehat{\mathbf{S}}_a$, and $\{\mathbf{F}^{(i)}\}_{i=1}^{L}$. The EMA teacher provides $\mathbf{g}_{\mathrm{full}}$, and the student is optimized by Eq. (2).

During fine-tuning, the pre-trained weights initialize the segmentation encoder-decoder, denoted by G_θ^{seg}. Given a supervised case $(\mathcal{C}, \mathbf{Y})$, the segmentation head Q_ϕ, with learnable parameters ϕ, converts the decoder representation into a probability map

$$\widehat{\mathbf{Y}} = Q_\phi\left(G_\theta^{\text{seg}}(\mathcal{C})\right) \in [0,1]^{K\times D\times H\times W}, \tag{16}$$

where K denotes the number of target classes. The segmentation objective is

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}, \tag{17}$$

where $\mathcal{L}_{\text{dice}}$ and $\mathcal{L}_{\text{ce}}$ denote the standard soft Dice and voxel-wise cross-entropy losses, respectively. For both BraTS19 and the carotid artery dataset, $\lambda_{\text{dice}} = 0.8$ and $\lambda_{\text{ce}} = 0.2$. The same objective is applied to three auxiliary decoder outputs with weights 0.5, 0.25, and 0.125 for deep supervision.

4 Experiments

4.1 Datasets

The Carotid Artery Dataset: The carotid artery dataset was provided by Guangdong Provincial People's Hospital and contains paired computed tomography (CT) and magnetic resonance imaging (MRI) volumes from 125 cases with carotid artery annotations, which were anonymized before analysis. Its use in this study was approved by the Ethics Committee of Guangdong Provincial People's Hospital (Approval No. KY2024-766-01), and the requirement for informed consent was waived. The study involved no additional patient interventions or risks and was conducted in accordance with the Declaration of Helsinki and applicable ethical regulations in China. The two modalities are matched at the case level, but have different spatial resolutions and are not jointly registered. The vessel foreground occupies only a small proportion of each volume and presents elongated, thin, and locally low-contrast structures. Therefore, this dataset is used to evaluate representation transfer in a small-structure segmentation setting. CT and MRI are treated as separate modalities in the modality-set formulation, and downstream segmentation is evaluated in both target modalities.

Brain Tumor Segmentation 2019 (BraTS19) Challenge [44,45]: BraTS19 provides multimodal MRI data for brain tumor segmentation, including T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) sequences. The sequences of each case are co-registered to a common anatomical space and share the same tumor annotation. Compared with the carotid artery dataset, BraTS19 contains larger but more heterogeneous targets, whose internal regions exhibit different appearances across MRI sequences. Each sequence is treated as one modality input. Segmentation performance is reported for the enhancing tumor (ET), tumor core (TC), and whole tumor (WT), thereby evaluating both the overall lesion extent and its nested internal regions.

For both datasets, the cases are divided into training and validation subsets using 80% of cases for training and 20% for validation. All quantitative results are reported on held-out validation cases.

4.2 Implementation Details

For both datasets, all volumes are resampled to an isotropic spacing of 1.5 mm × 1.5 mm × 1.5 mm. Linear interpolation is applied to images, whereas nearest-neighbor interpolation is used for labels. Each resampled volume is padded or cropped to 128^3, followed by modality-wise z-score normalization. The unified 128^3 input size is used for pre-training, fine-tuning, and evaluation on both datasets. During evaluation, ET, TC, and WT are derived from the original BraTS19 labels, while the carotid artery annotations are converted into binary vessel masks.

The pre-training augmentation includes random blurring, noise perturbation, cutout, flipping, and random crop-resize. Foreground-prioritized patch sampling is used during carotid fine-tuning to reduce the dominance of background-only crops. The backbone contains modality-specific stems, a shared encoder, a reconstruction decoder, and two Transformer encoder layers. Each modality-specific stem outputs a 32-channel feature map. The four subsequent shared encoder stages use 64, 128, 256, and 512 channels, respectively, and the reconstruction decoder outputs a 32-channel feature map. Each available modality contributes one modality-level token.

Pre-training follows the objectives and masking strategy described in Section 3. AdamW is used with an initial learning rate of 10^{-4}, weight decay of 10^{-4}, and batch size of 2. Gradient accumulation is performed over 2 iterations for BraTS19 and 3 iterations for the carotid artery dataset. The learning rate is controlled by linear warm-up followed by cosine annealing. The structure head and SL prediction blocks use the default PyTorch initialization. Convolutional weights use Kaiming-uniform initialization, and biases use fan-in-scaled uniform initialization. The affine instance-normalization parameters start at unit scale and zero offset. Early stopping monitors validation loss with a minimum improvement of 10^{-4}; training stops after 20 consecutive non-improving epochs during pre-training or 25 during fine-tuning. Downstream segmentation uses AdamW with an initial learning rate of 10^{-3}, weight decay of 10^{-4}, and batch size of 2.

4.3 Performance Comparison on BraTS19

Compared methods and evaluation setting: The compared methods cover convolutional networks, Transformer-based architectures, volumetric segmentation models, and self-supervised pre-training methods. Scratch fine-tuning and SLC use the same proposed backbone and supervised fine-tuning protocol; scratch starts from random initialization, whereas SLC starts from the proposed pre-training. This matched pair isolates the effect of SLC pre-training. SLC denotes the complete framework containing SSR, Hybrid Mask, SL, and S2F regularization.

On BraTS19, SegResNet, UNet, the UNet Transformer (UNETR) [2], EM-Net for efficient channel and frequency learning [46], the Large Kernel Vision Mamba UNet (LKM-UNet) [47], and the large-kernel volumetric ConvNet 3D UX-Net [48] are included as without-pretraining architecture baselines. Semantic Genesis [21], Swin UNETR [5], adaptive and hierarchical grid mask image modeling (GMIM) [33], volume contrastive learning (VoCo) [24], and OpenMind-MAE [7] represent pre-training-based methods for volumetric medical images. Dice measures regional overlap, whereas the 95th-percentile Hausdorff distance (HD95) measures boundary discrepancy while reducing sensitivity to isolated outliers.

Quantitative and qualitative results: Table 1 reports the quantitative comparison on BraTS19, grouped by whether pre-training is used. Against the matched scratch model, SLC increases ET, TC, and WT Dice from 84.09%, 88.46%, and 90.31% to 84.56%, 89.09%, and 91.38%, raising mean Dice from 87.62% to 88.34%. The gains cover both the overall lesion extent and the nested internal regions.

Statistical significance was assessed using two-sided paired Wilcoxon signed-rank tests with Holm correction across the three primary comparisons. The paired mean-Dice improvement on BraTS19 remained significant after Holm correction, with an adjusted p-value of 0.0359.

SLC also reduces mean HD95 from 3.792 to 3.498 mm. The largest HD95 gain appears on TC, where the value decreases from 3.854 to 3.307 mm, indicating more coherent internal tumor boundaries.

Among the without-pretraining baselines, UXNet obtains the strongest mean Dice of 88.04%, and EM-Net obtains the lowest mean HD95 of 3.561 mm. The matched scratch model reaches a mean Dice of 87.62% and a mean HD95 of 3.792 mm, providing the direct supervised reference for evaluating the effect of SLC pre-training under the same backbone.

Table 1: Comparison on the BraTS19 dataset. Bold and underlined values indicate the best and second-best results across all compared methods, respectively. ↑ and ↓ indicate that higher and lower values are better, respectively.

Method	Dice (%)↑				HD95 (mm)↓			
	ET	TC	WT	Mean	ET	TC	WT	Mean
Without pre-training								
SegResNet	75.33	82.66	86.41	81.47	16.82	19.11	16.18	17.37
UNETR [2]	81.73	83.52	89.80	85.02	7.104	11.68	11.48	10.09
UNet	83.12	87.88	90.72	87.24	2.354	3.982	5.798	4.045
EM-Net [46]	83.45	88.23	90.58	87.42	2.601	3.808	**4.274**	3.561
LKM-UNet [47]	**85.16**	87.74	91.15	88.02	2.242	4.363	4.513	3.706
UXNet [48]	84.56	88.22	91.31	88.04	2.262	**3.153**	5.561	3.659
Scratch fine-tuning (ours)	84.09	88.46	90.31	87.62	2.535	3.854	4.988	3.792
With pre-training								
Semantic Genesis [21]	82.87	84.84	88.40	85.37	5.716	9.855	15.97	10.51
Swin UNETR [5]	84.08	88.50	90.76	87.78	**2.207**	4.076	5.506	3.930
GMIM [33]	71.81	81.62	90.04	81.16	17.99	28.20	12.51	19.57
VoCo [24]	83.17	81.37	91.38	85.31	9.370	7.342	5.339	7.350
OpenMind-MAE [7]	84.13	85.31	**91.63**	87.02	9.259	5.603	4.712	6.525
SLC (ours)	84.56	**89.09**	91.38	**88.34**	2.431	3.307	4.756	**3.498**

Within the with-pretraining group, SLC achieves the highest mean Dice and the lowest mean HD95. Across all methods, LKM-UNet obtains the highest ET Dice, OpenMind-MAE obtains the highest WT Dice, and SLC obtains the highest TC Dice while ranking second on ET and WT. SLC also exceeds UXNet, the strongest without-pretraining mean-Dice baseline, by 0.30 percentage points. Compared with OpenMind-MAE, SLC improves mean Dice by 1.32 percentage points and reduces mean HD95 by 3.027 mm.

This pattern is consistent with the design of SLC. SSR supplies three-channel structural targets for masked reconstruction. SL uses scale-specific single-channel gradient targets to supervise transitions between neighboring decoder scales. Together, these objectives help preserve the broad lesion layout while retaining internal boundaries.

Fig. 3 presents qualitative results for tumors of different sizes. SLC produces more complete WT contours and more coherent TC and ET regions, especially in cases with thin tumor extensions or complex internal boundaries. For large tumors, most methods recover a similar overall extent, while SLC better preserves the nested WT, TC, and ET organization. This agrees with the quantitative gains in TC and WT Dice.

4.4 Performance Comparison on the Carotid Artery Dataset

Compared methods and evaluation setting: On the carotid artery dataset, SLC is compared with SegResNet, UNet, UNETR, Swin UNETR, Semantic Genesis, EM-Net, UXNet, U-Mamba, and the matched scratch control under the same case split, preprocessing, crop size, and checkpoint selection criterion. Dice measures regional overlap, whereas HD95 evaluates boundary discrepancy while reducing sensitivity to isolated outliers.

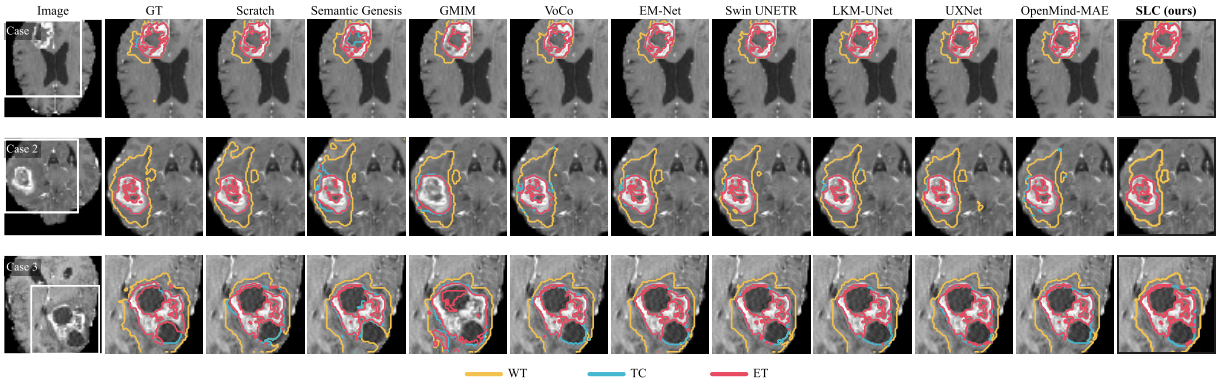

Figure 3: Qualitative comparison on BraTS19 across small, medium, and large tumor extents. Gold, cyan, and red contours denote WT, TC, and ET, respectively.

Quantitative results: Table 2 reports the segmentation results on the carotid artery dataset. Compared with the matched scratch control, SLC improves CT Dice from 89.24% to 91.07% and reduces CT HD95 from 2.439 to 1.560 mm. On MRI, Dice increases from 71.83% to 74.39%, while HD95 decreases from 4.576 to 4.080 mm. After Holm correction across the three primary comparisons, the adjusted p-values for the CT and MRI Dice improvements were both 0.0375.

Table 2: Comparison on the carotid artery dataset. Bold and underlined values indicate the best and second-best results across all compared methods, respectively. ↑ and ↓ indicate that higher and lower values are better, respectively.

Method	Dice (%)↑		HD95 (mm)↓	
	CT	MRI	CT	MRI
UNet	86.04	66.67	2.313	<u>4.521</u>
SegResNet	86.43	65.96	<u>1.705</u>	4.651
UNETR [2]	86.90	64.65	2.968	7.832
Swin UNETR [5]	88.19	70.86	2.401	4.783
Semantic Genesis [21]	88.96	67.94	2.125	5.208
EM-Net [46]	87.30	69.64	4.690	4.749
UXNet [48]	88.28	69.85	2.361	4.549
U-Mamba	<u>90.26</u>	**75.09**	1.721	4.525
Scratch fine-tuning	89.24	71.83	2.439	4.576
SLC (ours)	**91.07**	<u>74.39</u>	**1.560**	**4.080**

Among the architecture baselines, U-Mamba obtains the highest Dice on CT and MRI. SLC exceeds U-Mamba by 0.81 percentage points on CT Dice and further reduces CT HD95 to 1.560 mm. On MRI, SLC ranks second in Dice, 0.70 percentage points below U-Mamba, but obtains the lowest HD95 of 4.080 mm. The pre-training baseline Semantic Genesis obtains CT and MRI Dice of 88.96% and 67.94%, respectively.

All methods obtain lower Dice on MRI than on CT, indicating that MRI vessel segmentation is the more challenging target. SLC nevertheless improves the matched scratch model by 2.56 percentage points on MRI Dice, compared with a 1.83-point gain on CT, and reduces HD95 in both modalities.

The carotid artery is thin, sparse, and sensitive to local interruptions. The consistent improvement of Dice and HD95 supports the use of structure-oriented pre-training for this small-structure setting.

4.5 Ablation Study

Table 3 reports the component ablation results on BraTS19 and the carotid artery dataset. The scratch model is the supervised reference, and each remaining row removes one pre-training component. The full SLC configuration obtains the best TC and WT Dice on BraTS19, the best CT Dice on the carotid artery dataset, and the lowest ET, TC, and CT HD95.

Table 3: Ablation study on BraTS19 and the carotid artery dataset. Bold and underlined values indicate the best and second-best results across all configurations, respectively. ↑ and ↓ indicate that higher and lower values are better, respectively.

Configuration	Dice (%)↑					HD95 (mm)↓				
	BraTS19			Carotid Artery		BraTS19			Carotid Artery	
	ET	TC	WT	CT	MRI	ET	TC	WT	CT	MRI
Scratch fine-tuning	84.09	88.46	90.31	89.24	71.83	2.535	3.854	4.988	2.439	4.576
Without Hybrid Mask	**84.79**	88.84	<u>91.15</u>	90.15	74.31	<u>2.456</u>	3.638	<u>4.684</u>	1.923	**3.780**
Without SL Loss	83.71	<u>88.95</u>	90.95	<u>90.51</u>	<u>74.77</u>	2.685	3.789	**4.308**	2.079	3.829
Without SSR Loss	84.07	88.71	90.87	90.23	**75.71**	2.485	<u>3.477</u>	4.736	<u>1.705</u>	<u>3.813</u>
SLC (ours)	<u>84.56</u>	**89.09**	**91.38**	**91.07**	74.39	**2.431**	**3.307**	4.756	**1.560**	4.080

Hybrid Mask mainly affects targets with limited structural occupancy. Adding it increases TC and WT Dice from 88.84% and 91.15% to 89.09% and 91.38%, and improves carotid CT Dice from 90.15% to 91.07%. Although the variant without Hybrid Mask gives a slightly higher ET Dice and lower MRI HD95, the full setting performs better on TC, WT, and carotid CT.

Removing the SL loss removes explicit prediction between adjacent structural scales. Restoring it improves ET, TC, and WT Dice by 0.85, 0.14, and 0.43 percentage points and reduces ET and TC HD95 by 0.254 and 0.482 mm. On carotid CT, Dice increases from 90.51% to 91.07%, and HD95 decreases from 2.079 to 1.560 mm. The variant without the SL loss gives the lowest WT HD95 and a higher MRI Dice, showing that the effect remains metric-dependent.

SSR determines the reconstructed pre-training content. Restoring SSR improves ET, TC, and WT Dice by 0.49, 0.38, and 0.51 percentage points, increases carotid CT Dice by 0.84 percentage points, and reduces ET, TC, and CT HD95. The variant without SSR achieves the highest MRI Dice, whereas the full model performs better across the three BraTS regions and carotid CT.

Overall, Hybrid Mask allocates supervision to structural regions, SSR defines segmentation-relevant reconstruction targets, and SL supervises structural transitions between neighboring scales. Their combination gives the most balanced performance across lesion and vessel segmentation.

4.6 Efficiency Analysis

Table 4 compares model size, floating-point operations (FLOPs), inference latency in milliseconds (ms), peak allocated graphics processing unit (GPU) memory in megabytes (MB), and mean Dice on BraTS19 under the same input protocol.

SLC achieves the highest mean Dice with 52.11 million parameters and 610.00 billion FLOPs. Its model size and computational cost remain below those of several larger baselines, while its peak memory is comparatively high but lower than that of LKM-UNet. During downstream inference, only the segmentation network is deployed.

Table 4: Forward-only downstream inference efficiency on BraTS19. ↑ and ↓ indicate that higher and lower values are better, respectively.

Method	Fine-tuned Parameters (million)	FLOPs (billion)↓	Inference Time (ms)↓	Peak GPU Memory (MB)↓	Mean Dice (%)↑
SegResNet	4.70	125.90	13.77	406.2	81.47
UNETR [2]	139.57	471.56	43.38	1323.4	85.02
UNet	4.81	11.85	5.38	101.4	87.24
Semantic Genesis [21]	19.08	1660.00	54.93	2436.9	85.37
GMIM [33]	15.71	158.90	48.86	1300.8	81.16
VoCo [24]	72.77	674.29	67.85	1635.9	85.31
OpenMind-MAE [7]	102.35	2050.00	31.28	1511.6	87.02
EM-Net [46]	31.87	386.95	68.46	1001.4	87.42
Swin UNETR [5]	62.19	622.71	67.05	1598.7	87.78
LKM-UNet [47]	64.36	2420.00	91.06	2331.6	88.02
UXNet [48]	53.06	1180.00	70.78	1373.1	88.04
SLC (ours)	52.11	610.00	63.47	1952.0	88.34

Table 5 reports the cumulative costs of the pre-training components. The EMA teacher adds 53.72 million resident parameters with 0.41% iteration-time overhead. The SSR loss adds 594 head parameters and increases iteration time from 0.5370 to 0.8943 s. The SL loss adds 0.269 million parameters, with 7.74% higher iteration time and 12.63% higher peak memory than the preceding configuration. The complete pre-training system contains 107.71 million resident parameters. Table 4 profiles the fine-tuned segmentation model, which contains 52.11 million parameters after modality fusion and removal of pre-training-only outputs.

Table 5: Cumulative pre-training overhead of SLC on BraTS19. Added parameters are relative to the preceding row. "+" denotes the cumulative addition of the indicated loss and its training-only module(s) to the preceding setting; "–" marks the reference setting, for which no incremental parameter count is reported. ↓ indicates that a lower value is better.

Setting	Added Parameters (million)	Resident Parameters (million)	Iteration Time (s)↓	Peak GPU Memory (MB)↓
S2F without EMA	–	53.72	0.5348	17255.2
+EMA teacher	+53.72	107.44	0.5370	17453.8
+SSR Loss	+0.0006	107.44	0.8943	19195.6
+SL Loss	+0.269	107.71	0.9636	21620.7

4.7 Segmentation-Relevant Structural Supervision

SLC replaces intensity reconstruction with SSR and uses the fine-detail response to guide part of the masked supervision. In Table 3, removing the SSR loss weakens the full model on the three BraTS19 regions and on carotid CT.

Fig. 4 illustrates how the three SSR channels respond to different target morphologies. On the BraTS example, persistence and stable core provide continuous foreground and boundary emphasis, and fine detail retains visible local boundary responses. On the carotid examples, fine detail is concentrated along the

thin vessel structures. The dataset-level results in Table 6 support these observations. Stable core shows the strongest average enrichment for ET and TC, whereas fine detail is strongest for carotid CT and MRI. The channels therefore provide complementary structural cues across the two segmentation settings.

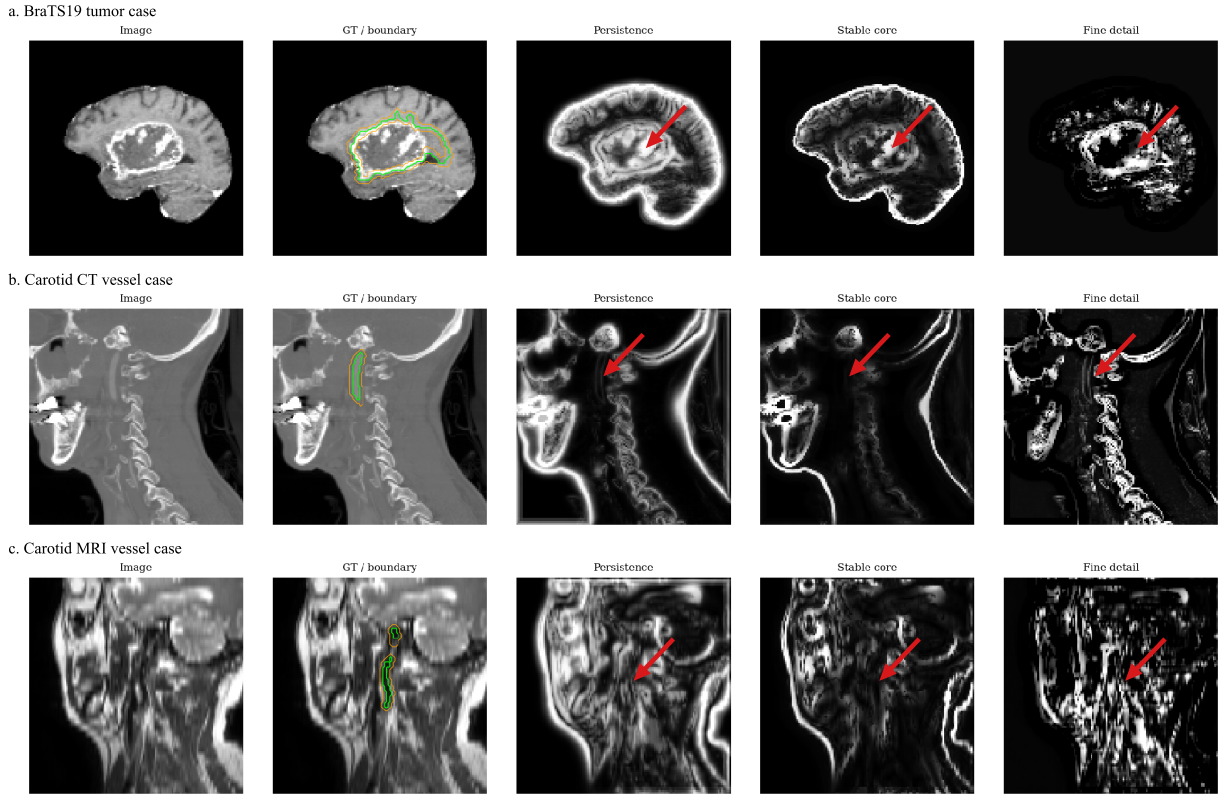

Figure 4: SSR structural target channels on BraTS19 and the carotid artery dataset. The rows show a BraTS tumor, a carotid CT vessel, and a carotid MRI vessel; the columns show the input image, annotated target, persistence response, stable-core response, and fine-detail response. Green and orange contours in the GT/boundary column denote the annotated foreground and boundary, respectively. Red arrows point to the corresponding target regions in the three structural-response maps. Persistence and stable core emphasize scale-stable responses, while fine detail emphasizes local high-frequency structures.

Table 6: SSR target response lift on foreground (FG) and boundary (BD) regions. Values above 1 indicate stronger response inside the target region than in its reference background. ↑ indicates that a higher value represents stronger target-region enrichment.

Dataset	Target	Mean Response Lift ↑					
		Persistence		Stable Core		Fine Detail	
		FG	BD	FG	BD	FG	BD
BraTS19	ET	1.92	1.98	2.13	2.19	1.89	1.65
BraTS19	TC	1.89	1.99	2.07	2.22	1.83	1.45
BraTS19	WT	1.63	1.84	1.74	2.03	1.29	1.34
Carotid CT	Vessel	0.62	0.59	0.46	0.53	2.00	1.55
Carotid MRI	Vessel	1.00	1.15	0.86	1.04	1.87	2.36

Table 6 reports foreground (FG) and boundary (BD) response lift. Let S_c, $c \in \{\text{pers}, \text{core}, \text{fine}\}$, denote a channel from Eq. (9). For foreground Ω, the boundary band is

$$B = \text{Dilate}_2(\Omega) \setminus \text{Erode}_2(\Omega), \tag{18}$$

where dilation and erosion use a 3D six-neighbor structuring element for two iterations. Boundary analysis complements region-level evaluation [32,49,50]. The response lifts are

$$L^c_{\text{FG}} = \frac{\text{mean}_{v \in \Omega} S_c(v)}{\text{mean}_{v \notin \Omega} S_c(v) + \varepsilon}, \tag{19}$$

$$L^c_{\text{BD}} = \frac{\text{mean}_{v \in B} S_c(v)}{\text{mean}_{v \notin B} S_c(v) + \varepsilon}, \tag{20}$$

where $\varepsilon = 10^{-8}$. Table 6 averages the sample-level lifts:

$$\overline{L}^c_r = \frac{1}{N} \sum_{n=1}^{N} L^c_{r,n}, \quad r \in \{\text{FG}, \text{BD}\}. \tag{21}$$

A lift above 1 indicates stronger structural response in the target region than in the corresponding reference region.

For ET and TC, the stable-core channel produces the strongest enrichment, with foreground/boundary lifts of 2.13/2.19 for ET and 2.07/2.22 for TC. Persistence also provides high responses in these regions. For WT, the same ordering remains, but the enrichment is weaker because WT covers a broader and more heterogeneous region.

The carotid artery shows a different pattern. For CT, the fine-detail response reaches a foreground/boundary lift of 2.00/1.55, whereas persistence and stable-core responses remain below 1.0. For MRI, the corresponding fine-detail lift is 1.87/2.36. This contrast supports the multi-channel design of SSR: stable responses are more informative for lesion-like regions, while high-frequency responses are more informative for thin vessels.

4.8 Neighboring-Scale Structural Transitions

Scale-aware representation learning also requires features to encode neighboring-scale structural transitions. Encoder-decoder networks produce feature pyramids, but multiple scales alone do not guarantee that these transitions are represented. SL therefore adds bidirectional prediction between neighboring structural scales.

Fig. 5 shows that the adjacent-scale prediction objectives decrease steadily on both datasets. To evaluate the learned representations beyond the training loss, a voxel-wise linear probe was trained on frozen encoder features to predict neighboring-scale structural targets. Table 7 reports mean absolute error (MAE), boundary MAE, and correlation across the eight directed predictions, while Fig. 6 presents the corresponding qualitative comparison.

Across BraTS19, carotid CT, and carotid MRI, SLC reduces overall MAE by 15.12%–20.98% and boundary MAE by 15.10%–21.68%, while increasing correlation by 16.59%–20.55%. The improvements are consistent across all eight prediction directions.

Fig. 6 compares the same frozen probe for Without SL Loss and SLC using an identical case, slice, and display range. For both $S^2 \to S^3$ and $S^3 \to S^2$, the SLC predictions more closely follow the neighboring-scale targets and produce smaller absolute errors in the principal high-response regions.

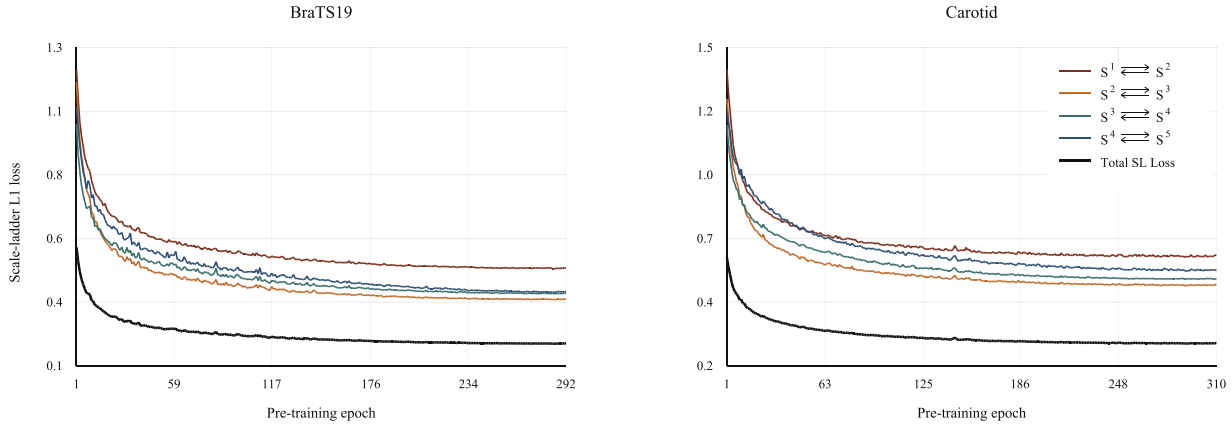

Figure 5: SL loss curves during pre-training. Colored curves denote bidirectional prediction losses between adjacent decoder scales, and the black curve denotes the total SL loss recorded by the training objective.

Table 7: Frozen-feature evaluation of neighboring-scale structural prediction. Bold values indicate the better result between the two pre-training settings for each dataset. ↑ and ↓ indicate that higher and lower values are better, respectively.

Dataset	Pre-Training	MAE↓	Boundary MAE↓	Correlation↑
BraTS19	Without SL Loss	0.4375	0.4986	0.5865
BraTS19	SLC (ours)	**0.3714**	**0.4233**	**0.6940**
Carotid CT	Without SL Loss	0.4057	0.3478	0.6943
Carotid CT	SLC (ours)	**0.3206**	**0.2724**	**0.8095**
Carotid MRI	Without SL Loss	0.4361	0.5803	0.6599
Carotid MRI	SLC (ours)	**0.3455**	**0.4827**	**0.7955**

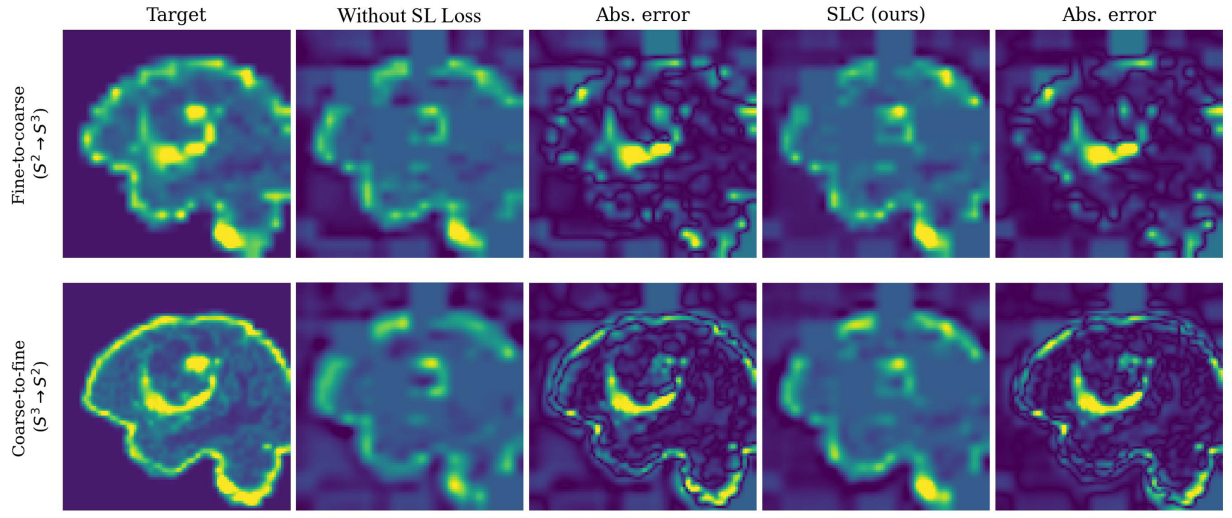

Figure 6: Matched frozen-probe visualization on BraTS19 for fine-to-coarse ($S^2 \to S^3$) and coarse-to-fine ($S^3 \to S^2$) prediction. Columns show the target, predictions, and absolute errors without the SL loss and with SLC using the same case, slice, and display range.

The two directions provide different constraints. Fine-to-coarse prediction encourages fine representations to retain the dominant anatomical layout after scale reduction, whereas coarse-to-fine prediction preserves information needed to recover local detail. This is consistent with Table 3, where removing the SL loss weakens several boundary-sensitive metrics, including carotid CT Dice and HD95.

5 Applicability and Limitations

SLC constructs its pre-training targets from image-derived scale-space responses, allowing the same objective to be configured for other segmentation tasks with multi-scale anatomical structure. In abdominal organ segmentation, persistence and stable-core responses may emphasize organ interfaces that remain visible across scales. In pulmonary vessel segmentation, fine-detail responses and adjacent-scale prediction may support thin branches and local continuity. These applications require task-specific selection of smoothing radii and masking parameters, followed by downstream validation.

The applicability of SLC depends on an encoder-decoder that exposes multi-scale decoder representations. Architectures with only a single output scale would require an additional feature hierarchy before adjacent-scale prediction can be applied. The SSR head, eight bidirectional SL prediction blocks, and full-view EMA teacher increase training time and memory during pre-training. The prediction heads and teacher branch are discarded before downstream inference.

The gradient-derived structural targets also depend on image resolution and quality. Extremely low resolution may erase thin structures before scale-space target construction, while strong noise can introduce spurious high-frequency responses. Future work will investigate resolution-adaptive scales and denoising or confidence-weighted targets for challenging image quality.

6 Conclusion

This study presented SLC, a structure-aware self-supervised framework for multimodal 3D medical image segmentation. SLC addresses the mismatch between intensity recovery and segmentation-relevant structural supervision and the lack of explicit supervision for neighboring-scale transitions. SSR and Hybrid Mask redirect pre-training toward sparse anatomical structures, SL learns cross-scale structural change through bidirectional prediction, and S2F stabilizes case-level representations using an auxiliary full-view reference.

Experiments on BraTS19 and the carotid artery dataset show that SLC improves over matched scratch fine-tuning on both tumor and vessel segmentation. The comparisons, ablations, and structural analyses collectively support the use of segmentation-relevant structural targets and explicit neighboring-scale prediction in multimodal 3D medical representation learning.

The current study is limited by its reliance on multi-scale decoder features and additional pre-training computation. Future work will improve structural target construction for low-resolution and noisy images through resolution-adaptive scales and denoising or confidence-weighted targets.

Acknowledgement: Not applicable.

Funding Statement: This work is supported by the National Natural Science Foundation of China under Grant 62273155, Science and Technology Projects in Guangzhou (2025B01J3018), Science and Technology Project of Ganzhou (2023LNS27051).

Author Contributions: Weiqing Liu: Conceptualization, Methodology, Writing original draft, Software, Review & editing. Bin Li: Conceptualization, Funding acquisition, Investigation, Methodology, Writing—review & editing,

Project administration, Validation. Lianfang Tian: Methodology. Qianhui Qiu: Data curation, Investigation, Resources, Validation. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The Brain Tumor Segmentation 2019 (BraTS19) dataset is publicly available through the BraTS challenge under its data access conditions [44,45]. The carotid artery dataset is not publicly available because patient-privacy requirements and institutional data-use restrictions prohibit external distribution.

Ethics Approval: The carotid artery dataset consists of retrospectively collected imaging data that were anonymized before analysis. Its use in this study was approved by the Ethics Committee of Guangdong Provincial People's Hospital (Approval No. KY2024-766-01), and the requirement for informed consent was waived. The study involved no additional patient interventions or risks and was conducted in accordance with the Declaration of Helsinki and applicable ethical regulations in China.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.
2. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: transformers for 3D medical image segmentation. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3–8; Waikoloa, HI, USA. p. 1748–58. doi:10.1109/wacv51458.2022.00181.
3. Roy S, Koehler G, Ulrich C, Baumgartner M, Petersen J, Isensee F, et al. MedNeXt: transformer-driven scaling of ConvNets for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2023. Cham, Switzerland: Springer Nature; 2023. p. 405–15. doi:10.1007/978-3-031-43901-8_39.
4. He K, Chen X, Xie S, Li Y, Dollar P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 15979–88. doi:10.1109/cvpr52688.2022.01553.
5. Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, et al. Self-supervised pre-training of swin transformers for 3D medical image analysis. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 20698–708. doi:10.1109/cvpr52688.2022.02007.
6. Chen Z, Agarwal D, Aggarwal K, Safta W, Balan MM, Brown K. Masked image modeling advances 3D medical image analysis. In: Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2023 Jan 2–7; Waikoloa, HI, USA. p. 1969–79. doi:10.1109/wacv56688.2023.00201.
7. Wald T, Ulrich C, Suprijadi J, Ziegler S, Nohel M, Peretzke R, et al. An OpenMind for 3D medical vision self-supervised learning. In: Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV); 2025 Oct 19–25; Honolulu, HI, USA. p. 23839–79. doi:10.1109/iccv51701.2025.02213.
8. Pan T, Tan Z, Guo K, Xu D, Xu W, Jiang C, et al. Structure-aware semantic discrepancy and consistency for 3D medical image self-supervised learning. In: Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision (ICCV); 2025 Oct 19–25; Honolulu, HI, USA. p. 20257–67. doi:10.1109/iccv51701.2025.01884.
9. Gao Y, Zhang Y, Wang C, Liu J, Varma M, Delbrouck JB, et al. Learning generalizable 3D medical image representations from mask-guided self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2026 Jun 3–7; Denver, CO, USA. p. 13744–54.
10. Tu J, Qin G, Zhao TZ, Valanarasu JMJ, Zhang S, Naumann T, et al. Masked-diffusion autoencoders for 3D medical vision representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2026 Jun 3–7; Denver, CO, USA. p. 22804–15.
11. Shit S, Paetzold JC, Sekuboyina A, Ezhov I, Unger A, Zhylka A, et al. clDice—a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 16555–64. doi:10.1109/cvpr46437.2021.01629.

12. Yang Z, Farsiu S. Directional connectivity-based segmentation of medical images. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 11525–35. doi:10.1109/cvpr52729.2023.01109.
13. Wang Y, Li Z, Mei J, Wei Z, Liu L, Wang C, et al. SwinMM: masked multi-view with swin transformers for 3D medical image segmentation. In: Medical image computing and computer assisted intervention—MICCAI 2023. Cham, Switzerland: Springer Nature; 2023. p. 486–96. doi:10.1007/978-3-031-43898-1_47.
14. Xing Z, Zhu L, Yu L, Xing Z, Wan L. Hybrid masked image modeling for 3D medical image segmentation. IEEE J Biomed Health Inform. 2024;28(4):2115–25. doi:10.1109/JBHI.2024.3360239.
15. Wang X, Wang R, Tian B, Zhang J, Zhang S, Chen J, et al. MPS-AMS: masked patches selection and adaptive masking strategy based self-supervised medical image segmentation. In: Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2023 Jun 4–10; Rhodes Island, Greece. p. 1–5. doi:10.1109/icassp49357.2023.10094657.
16. Wang X, Wang R, Lukasiewicz T, Xu Z. AMLP: adjustable masking lesion patches for self-supervised medical image segmentation. IEEE Trans Med Imaging. 2026;45(5):1730–46. doi:10.1109/TMI.2025.3636922.
17. Zhou HY, Guo J, Zhang Y, Han X, Yu L, Wang L, et al. nnFormer: volumetric medical image segmentation via a 3D transformer. IEEE Trans Image Process. 2023;32:4036–45. doi:10.1109/TIP.2023.3293771.
18. Zhuang J, Wu L, Wang Q, Fei P, Vardhanabhuti V, Luo L, et al. MiM: mask in mask self-supervised pre-training for 3D medical image analysis. IEEE Trans Med Imaging. 2025;44(9):3727–40. doi:10.1109/TMI.2025.3564382.
19. Yang S, Wang H, Xing Z, Chen S, Yang Q, Mao Y, et al. SegDINO: introducing multi-scale structure into DINO for efficient medical image segmentation. arXiv:2606.17972. 2026.
20. Zhou Z, Sodha V, Pang J, Gotway MB, Liang J. Models genesis. Med Image Anal. 2021;67(4):101840. doi:10.1016/j.media.2020.101840.
21. Haghighi F, Hosseinzadeh Taher MR, Zhou Z, Gotway MB, Liang J. Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration. Med Image Comput Comput Assist Interv. 2020;12261(6):137–47. doi:10.1007/978-3-030-59710-8_14.
22. Chaitanya K, Erdil E, Karani N, Konukoglu E. Contrastive learning of global and local features for medical image segmentation with limited annotations. Adv Neural Inf Process Syst. 2020;33:12546–58. doi: 10.3929/ethz-b-000443425.
23. He Y, Yang G, Ge R, Chen Y, Coatrieux JL, Wang B, et al. Geometric visual similarity learning in 3D medical image self-supervised pre-training. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 9538–47. doi:10.1109/cvpr52729.2023.00920.
24. Wu L, Zhuang J, Chen H. VoCo: a simple-yet-effective volume contrastive learning framework for 3D medical image analysis. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 22873–82. doi:10.1109/cvpr52733.2024.02158.
25. Xie Y, Zhang J, Xia Y, Wu Q. UniMiSS: universal medical self-supervised learning via Breaking dimensionality barrier. In: Computer Vision—ECCV 2022. Cham, Switzerland: Springer Nature; 2022. p. 558–75. doi:10.1007/978-3-031-19803-8_33.
26. Gao F, Wang S, Zhang F, Zhou HY, Wang Y, Wang C, et al. Cross-dimensional medical self-supervised representation learning based on a Pseudo-3D transformation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 178–88. doi:10.1007/978-3-031-72120-5_17.
27. Haghighi F, Hosseinzadeh Taher MR, Gotway MB, Liang J. Self-supervised learning for medical image analysis: discriminative, restorative, or adversarial? Med Image Anal. 2024;94(2):103086. doi:10.1016/j.media.2024.103086.
28. Han J, Wang F, Shen X, Cao F. MedCSS: a causal self-supervised approach for hierarchical feature consistency in 3D medical imaging. Front Neurosci. 2026;20:1739716. doi:10.3389/fnins.2026.1739716.
29. Gupta S, Hu X, Kaan J, Jin M, Mpoy M, Chung K, et al. Learning topological interactions for multi-class medical image segmentation. In: Computer Vision—ECCV 2022. Cham, Switzerland: Springer Nature; 2022. p. 701–18. doi:10.1007/978-3-031-19818-2_40.

30. Ju J, Qiu D, Lei H, Ren S, Zhao W, Xu P, et al. CDI-NSTSEG: a clinical diagnosis-inspired effective and efficient framework for non-salient small tumor segmentation. IEEE J Biomed Health Inform. 2024;28(12):7469–79. doi:10.1109/JBHI.2024.3440925.
31. Ju J, Zhang T, Song W, Xiao Z, Tu H, Guan Z, et al. Collaborative learning of dynamic and static information for medical image segmentation. Inf Fusion. 2026;125(1):103509. doi:10.1016/j.inffus.2025.103509.
32. Liu C, Cheng Y, Tamura S. Masked image modeling-based boundary reconstruction for 3D medical image segmentation. Comput Biol Med. 2023;166(Pt 3):107526. doi:10.1016/j.compbiomed.2023.107526.
33. Qi L, Jiang Z, Shi W, Qu F, Feng G. GMIM: self-supervised pre-training for 3D medical image segmentation with adaptive and hierarchical masked image modeling. Comput Biol Med. 2024;176(1):108547. doi:10.1016/j.compbiomed.2024.108547.
34. Pissas T, Ravasio CS, Da Cruz L, Bergeles C. Multi-scale and cross-scale contrastive learning for semantic segmentation. In: Computer Vision—ECCV 2022. Cham, Switzerland: Springer Nature; 2022. p. 413–29. doi:10.1007/978-3-031-19818-2_24.
35. Zeng Z, Xiao J, Yi S, Liu Q, Zhu Y. M3-TransUNet: medical image segmentation based on spatial prior attention and multi-scale gating. J Imaging. 2025;12(1):15. doi:10.3390/jimaging12010015.
36. Lin Q, Li G, Pan X, Lin Y, Nabi FG, Li S, et al. SDNAL-Seg: multi-scale selective downsampling and non-adjacent layers guidance for medical image segmentation. Signal Image Video Process. 2026;20(7):408. doi:10.1007/s11760-026-05376-5.
37. Singh AK, Ranjan A, Prusty MR, Katal N. CA2PNet: a context-aware multi-scale architecture with adaptive attention and progressive dilated convolutions for biomedical image segmentation. Front Artif Intell. 2026;9:1802033. doi:10.3389/frai.2026.1802033.
38. Zhang Y, He N, Yang J, Li Y, Wei D, Huang Y, et al. mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2022. Cham, Switzerland: Springer Nature; 2022. p. 107–17. doi:10.1007/978-3-031-16443-9_11.
39. Liu H, Wei D, Lu D, Sun J, Wang L, Zheng Y. M3AE: multimodal representation learning for brain tumor segmentation with missing modalities. Proc AAAI Conf Artif Intell. 2023;37(2):1657–65. doi:10.1609/aaai.v37i2.25253.
40. Zeng Z, Peng Z, Yang X, Shen W. Missing as masking: arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 424–33. doi:10.1007/978-3-031-72111-3_40.
41. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 9726–35. doi:10.1109/cvpr42600.2020.00975.
42. Grill JB, Strub F, Altche F, Tallec C, Richemond PH, Buchatskaya E, et al. Bootstrap your own latent: a new approach to self-supervised learning. Adv Neural Inf Process Syst. 2020;33:21271–84.
43. Zhou X, Sun Y, Deng M, Chu WCW, Dou Q. Robust semi-supervised multimodal medical image segmentation via cross modality collaboration. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 57–67. doi:10.1007/978-3-031-72378-0_6.
44. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data. 2017;4(1):170117. doi:10.1038/sdata.2017.117.
45. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993–2024. doi:10.1109/TMI.2014.2377694.
46. Chang A, Zeng J, Huang R, Ni D. EM-Net: efficient channel and frequency learning with mamba for 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 266–75. doi:10.1007/978-3-031-72114-4_26.

47. Wang J, Chen J, Chen D, Wu J. LKM-UNet: large kernel vision mamba UNet for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 360–70. doi:10.1007/978-3-031-72111-3_34.
48. Lee HH, Bao S, Huo Y, Landman BA. 3D UX-Net: a large kernel volumetric ConvNet modernizing hierarchical transformer for medical image segmentation. In: Proceedings of the International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda.
49. Kervadec H, Bouchtiba J, Desrosiers C, Granger E, Dolz J, Ben Ayed I. Boundary loss for highly unbalanced segmentation. Med Image Anal. 2021;67(2):101851. doi:10.1016/j.media.2020.101851.
50. Cheng B, Girshick R, Dollar P, Berg AC, Kirillov A. Boundary IoU: improving object-centric image segmentation evaluation. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 15329–37. doi:10.1109/cvpr46437.2021.01508.