Towards Voxel Spacing Consistency for Medical Image Segmentation

Xin You, Runze Yang, Minghui Zhang, Hanxiao Zhang, Han Li, Yi Yu, Jie Yang, *Senior Member, IEEE*, Nassir Navab, *Fellow, IEEE*, and Yun Gu, *Member, IEEE*

***Abstract*—Volumetric medical image segmentation is essential for both preoperative diagnosis and intraoperative guidance. While recent years have witnessed rapid progress in segmentation architectures, comparatively little attention is paid to the physical voxel spacing of anatomical data. Indeed, volumetric image resampling is a ubiquitous preprocessing step before segmentation, yet its interaction with downstream segmentation has not been systematically exploited. In this work, we study the correlation between image resampling and segmentation, and propose Consispace, a semantic-aware resampling framework that achieves consistent voxel spacing in the axial direction while preserving anatomical and semantic consistency. Consispace introduces an ODE-based anatomical constraint to model inter-slice dynamics with a continuous interpolator, enabling faithful reconstruction under complex anatomical transitions beyond discrete interpolation. To further couple resampling with segmentation objectives, we leverage dense features from a pretrained vision model to build intra-slice semantic correlation maps and inject class-wise semantic consistency via feature reweighting during resampling. Both intra-slice and inter-slice constraints are integrated into an implicit neural network, supporting arbitrary-scale resampling. Extensive experiments on multiple datasets demonstrate that Consispace achieves superior reconstruction quality and perceptual fidelity, produces smoother inter-slice anatomy, and improves downstream segmentation performance when used as a preprocessing step. Codes are available at https://github.com/AlexYouXin/Consispace.**

***Index Terms*—Volumetric image resampling, Spacing consistency, ODE, Medical image segmentation.**

I. Introduction

Volumetric medical image segmentation plays a significant role in numerous clinical applications [1], including preoperative diagnostic assistance [2], treatment planning [3], intra-operative guidance [4], and longitudinal assessment of tumor progression [5]. To achieve precise and robust segmentation for anatomical structures, existing approaches predominantly emphasize increasingly sophisticated network architectures [1], [6]–[8] and anatomically guided loss functions [9]–[12]. In addition, recent foundation models, such as MedSAM2 [13] and SAM-Med3D [14], have sought to investigate domain transfer from natural images to the medical imaging domain.

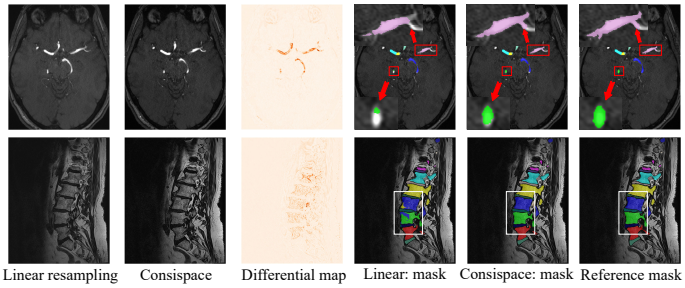

Fig. 1: The performance comparison between linear interpolation and the proposed Consispace on medical image resampling and segmentation. Differential map corresponds to the subtraction of resampled images by linear interpolation and Consispace.

However, comparatively little attention has been paid to intrinsic imaging properties, such as the physical voxel spacing of anatomical data, especially in the axial direction.

A recent study, Hyperspace [15], presents a quantitative analysis of how axial voxel spacing influences the performance of medical image segmentation models. Specifically, the most straightforward solution is to deal with images at their native resolution. However, Hyperspace claims that unifying the slice spacing for training and testing datasets can alleviate the inconsistency of physical spacing, thereby enhancing models' consistent representation of anatomical structures. Notably, this finding is in line with the design choice adopted by state-of-the-art frameworks such as nnU-Net [1] and TotalSegmentator [16]. Specifically, nnU-Net resamples data linearly to the median voxel spacing, to realize spacing consistency across all samples. Following this practice, almost all approaches adopt the standard linear interpolation strategy to achieve data harmonization during the preprocessing stage.

However, this naturally raises an important question: "**Is the current practice of linear resampling truly the optimal strategy?**" Medical image resampling couples the voxel spacing in the real world with discrete image resolutions, but it fundamentally performs voxel-wise downsampling and upsampling on volumetric data [1]. Downsampling is less problematic, While upsampling constitutes an ill-posed inverse problem—inferring missing information from limited observations. Consequently, linear interpolation cannot perfectly model complicated inter-slice shape and texture variations, exhibiting two limitations. Firstly, linear interpolation may cause anatomical non-smoothness along the axial direction, leading to structural discontinuities (Fig.1: the 1^{st} row). Secondly, it can result in distorted boundaries between different structures, violating semantic consistency across class-wise

This work was supported in part by National Key R&D Program of China (2022ZD0212400), Shanghai Municipal Commission of Economy and Informatization (2025-GZL-RGZN-BTBX-02023) and Suzhou Industrial Park Oriental Huaxia Cardiovascular Health Research Institute (GUSU2023002). (Corresponding authors: Yun Gu; Yi Yu.)

X. You, R. Yang, M. Zhang, H. Zhang, J. Yang and Y. Gu are with the Institute of Medical Robotics, Shanghai Jiao Tong University, 200240 Shanghai, China (Email: {sjtu_youxin, runze.y, minghuizhang, hanxiao.zhang, jieyang}@sjtu.edu.cn, yungu@ieee.org).

Y. Yu is with the Department of Ultrasound, Shanghai Chest Hospital, 200052 Shanghai, China (Email: yuyiyy2021@163.com).

H. Li and N. Navab are with CAMPAR, Technical University of Munich, 80333 Munich, Germany (Email: {tum_han.li, nassir.navab}@tum.de).

anatomies (Fig.1: the 2^{nd} row) [17]. Furthermore, upsampling in this work focuses on interpolation along the axial direction. Essentially, it is analogous to volumetric super-resolution (VSR) [18], but differs from conventional 2D super-resolution tasks [19]. However, prior VSR models [18], [20]–[23] treat resampling as an upstream task independent of downstream segmentation, thereby ignoring the intrinsic task correlation.

In this paper, we pioneer the exploration of the correlation between volumetric image resampling and medical image segmentation. On the one hand, to boost the segmentation performance on volumetric images, an effective image resampling framework is designed to unify the physical spacing distribution. Specifically, we propose an ordinary differential equation (ODE)-based anatomical constraint in the resampling network, to characterize the inter-slice spatial dynamics through an ODE-based interpolator. This interpolator enables the network to describe complex anatomical transitions in a continuous and differentiable manner, thereby mitigating the limitations of conventional discrete modeling approaches. Meanwhile, it will boost segmentation through the unified voxel spacing, as demonstrated by the aforementioned study [1], [15].

On the other hand, segmentation guidance is effective in promoting the process of medical image resampling. Concretely, we introduce semantic consistency into class-wise anatomies. On the ground that pretrained vision models provide high-quality dense features, semantic correlation maps can be extracted to encode contextual relations within each slice. Further, the boundaries of class-wise anatomies are clearly delineated semantically, which will enhance segmentation performance in the downstream task. Both intra-slice and inter-slice constraints are embedded into an implicit neural representation (INR) framework [19], [20] to achieve image resampling at random scales.

The proposed model naturally generalizes on multiple datasets by unifying data with heterogeneous voxel spacings into a Consistent voxel space. Thus, we nominate this resampling framework as **Consispace**. Experiments demonstrate that Consispace achieves promising image reconstruction performance and feature-level visual perceptual effects. Qualitative results illustrate that Consispace reveals resampled volumetric images with better inter-slice anatomical smoothness. Besides, in the downstream segmentation evaluation based on nnUNet [1], Swin UNETR [6] and recent MedSAM-2 [24], datasets preprocessed by Consispace bring more significant increases compared to standard linear resampling. Our main contributions are summarized as follows:

- We present the first systematic investigation of the bi-directional interaction between volumetric data resampling and medical image segmentation.
- Firstly, an INR-based resampling pipeline, Consispace, is introduced to harmonize the physical spacing distribution, which will strengthen segmentation performance in the downstream task. To achieve uniform physical spacing, we introduce an ODE-based anatomical constraint in the resampling network to boost inter-slice smoothness through a continuous formulation.
- Secondly, class-wise semantic consistency is enforced to improve medical image resampling. Intra-slice semantic correlation maps are derived from a pretrained vision model to produce semantically coherent anatomical structures.
- Extensive results show that Consispace achieves superior resampling quality, yields inter-slice anatomical continuity, and consistently improves downstream segmentation performance when used as a preprocessing step.

II. Related Work

A. Medical Image Segmentation

U-shaped encoder–decoder architectures [25] have dominated this field. Firstly, convolutional neural networks (CNNs) [1], [7], [26], [27] are fully leveraged due to the inductive bias and locality properties of convolutional kernels. In particular, 3D U-Net variants and strong engineering-oriented baselines such as nnU-Net [1] have demonstrated robust performance across diverse anatomical structures, largely due to their effective multi-scale feature aggregation and carefully designed training pipelines. Further, Transformer-based models [6], [28]–[31] have been introduced to enhance global context modeling in 3D volumes, typically by combining self-attention encoders [32] with U-shaped decoders. Concretely, Swin UNETR [6] introduced the windowed attention mechanism to alleviate the cubic complexity in volumetric settings. Since self-attention incurs quadratic computational costs with respect to the token number, Mamba-based networks [33]–[37] have been proposed to model the global context within medical images with linear-time complexity. More recently, foundation-model paradigms represented by Segment Anything Model (SAM) [38] have shifted attention toward promptable segmentation and cross-domain generalization; medical adaptations [13], [24], [39], [40] such as SAM-Med3D [14] aim to transfer knowledge from large-scale natural-image pretraining to medical imaging tasks. Besides, DINOv3 [41] has shown promising performance in medical image segmentation due to enriched deep features. Despite advances in architectural design and pretraining strategies, the impact of dataset-specific properties on model training remains under-explored, particularly voxel spacing and resampling protocols.

B. Volumetric Image Super-resolution

This task is equivalent to medical image upsampling, which is much more challenging compared with image downsampling. Specifically, Peng et. al pioneeringly introduce a Spatial-Aware Interpolation NeTwork (SAINT) [18], which utilizes voxel spacing information to provide desirable details. TVSRN [21] adopts a pure Transformer network for volumetric super-resolution, achieving a better trade-off between image quality, the number of parameters, and inference time. ArSSR [20] incorporates the implicit voxel function via deep neural networks, to realize the arbitrary upsampling-rate reconstruction of high-resolution images from any low-resolution images. CycleINR [22] further enhances the grid sampling with local attention and mitigates over-smoothing by integrating the cycle consistent loss. DP-INR [23] effectively combines the prior information learning capabilities of diffusion models [42] with the global information integration capacity of

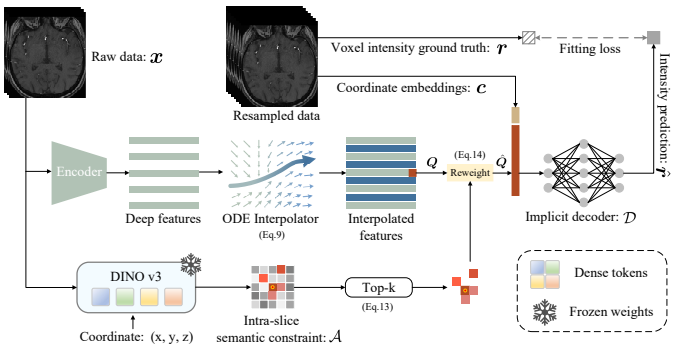

Fig. 2: The overall pipeline of Consispace. The ODE-based interpolator models inter-slice dynamics of anatomical structures in the axial direction. Besides, a pretrained vision model is introduced to extract intra-slice contextual correlations, which are incorporated to reweight interpolated features. Finally, an implicit neural network is proposed to aggregate both inter-slice and intra-slice anatomical and semantic constraints.

spatio-temporal implicit attention. More recently, arbitrary-scale paradigms such as AnySR [43] extend reconstruction to any-scale settings, while progressive implicit refinement for anisotropic MRI [44] models continuous resolution transitions. JOANet [45] further couples super-resolution pre-processing with segmentation via joint optimization, yet operates on 2D slices without addressing volumetric voxel-spacing consistency. However, the aforementioned methods fail to explicitly model the inter-slice correlation, leading to anatomical non-smoothness along the inter-slice direction and causing structural discontinuities.

C. Spacing-oriented Segmentation

SynthSeg [46] is the first segmentation CNN robust against changes in contrast and resolution. SynthSeg is trained with synthetic data sampled from a generative model conditioned on segmentations. Crucially, a domain randomisation strategy is adopted by randomising the contrast and resolution of the synthetic training data. Consequently, SynthSeg can segment real scans from a wide range of target domains without retraining or fine-tuning. HyperSpace [15] proposes to condition segmentation models on the voxel spacing using hypernetworks. This approach allows processing images at their native resolutions or at resolutions adjusted to the hardware and time constraints at inference time. However, these two methods do not consider the task interaction between image segmentation and medical image resampling.

III. Methodology

A. Problem Formulation and Whole Pipeline

The ultimate objective is to achieve promising segmentation performance on volumetric datasets with diverse voxel spacing. To achieve that, we need to optimize the following objective function:

$$\tilde{\mathcal{S}} = \arg\min_{\theta} \mathcal{L}_s(\mathcal{S}_\theta(\boldsymbol{X}), \boldsymbol{Y}), \tag{1}$$

where θ means the parameters of segmentation networks $\mathcal{S}_\theta$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ refer to the image and mask sets. $\mathcal{L}_s$ corresponds to the segmentation loss function, which provides the optimal solution $\tilde{\mathcal{S}}$ in the parameter space. However, as Hyperspace [15] and other literature [1], [47], [48] claim, unified voxel spacing can alleviate the inconsistency of physical spacing, thereby boosting anatomical representations of segmentation models. Thus, we design an image resampling framework to enforce a harmonized voxel spacing before feeding the training and testing datasets to the segmentation network. The specific target is as follows:

$$\mathcal{R}^* = \arg\min_{\phi} \mathcal{L}_r(\mathcal{R}_\phi(\boldsymbol{X}_l), \boldsymbol{X}_h), \tag{2}$$

where ϕ means the parameters of resampling networks $\mathcal{R}_\phi$. Since the downsampling operation is generally less challenging, we mainly focus on the upsampling operation. Thus, low-resolution and high-resolution image pairs $\boldsymbol{X}_l$, $\boldsymbol{X}_h$ are constructed to model the upsampling process. The reconstruction loss $\mathcal{L}_r$ helps generate the optimal reconstruction network $\mathcal{R}^*$. Therefore, the final segmentation network $\mathcal{S}^*$ should satisfy:

$$\mathcal{S}^* = \arg\min_{\theta} \mathcal{L}_s(\mathcal{S}_\theta(\mathcal{R}^*(\boldsymbol{X})), \boldsymbol{Y}). \tag{3}$$

In this work, any U-shaped encoder-decoder architecture can be adopted to represent $\mathcal{S}^*$. While for the medical image resampling, we choose an implicit neural network [20] to achieve image downsampling and upsampling in random scales. This pipeline is aimed at Consistent voxel spacing, termed **Consispace**. An overview of Consispace is shown in Fig.2.

B. ODE-based Inter-slice Dynamics

In this work, volumetric data is regarded as an ordered sequence of 2D axial slices, $\boldsymbol{x} = [\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(D')}]$. Each slice $\boldsymbol{x}^{(i)}$ belongs to $\mathbb{R}^{H\times W}$. Since adjacent slices within a volume typically exhibit gradual anatomical variations, we treat the slice index i as a pseudo-time variable to model inter-slice transitions. This perspective allows anatomical representations to evolve in a continuous latent space, providing an anatomical regularization mechanism for volumetric resampling beyond voxel-wise interpolation. Specifically, we assume that the target 3D anatomy can be represented by a time-dependent latent field $\boldsymbol{Q}(t) \in \mathbb{R}^{D\times H\times W\times d}$, whose evolution is governed by an ODE-inspired dynamics [49]:

$$\frac{\partial \boldsymbol{Q}}{\partial t} = \boldsymbol{v}(\boldsymbol{Q}, t). \tag{4}$$

Here, $\boldsymbol{Q}(t)$ encodes slice-wise anatomical representations, and the time-varying velocity field $\boldsymbol{v}$ characterizes their smooth evolution along the axial direction.

The continuous formulation above serves as a modeling prior for inter-slice anatomical evolution, while in practice the computation is performed at discrete slice index. Starting from an initial condition $\boldsymbol{Q}^{(0)}$, we discretize the trajectory of $\boldsymbol{Q}(t)$ with a step size Δt, obtaining latent states $\boldsymbol{Q}^{(i\Delta t)}$ for $i = 1, \ldots, D$. The dynamics in Eq. (4) are then numerically approximated by the Euler method:

$$\boldsymbol{Q}^{(i)} = \boldsymbol{Q}^{(i-1)} + \Delta t\, \boldsymbol{v}^{(i-1)}\left(\boldsymbol{Q}^{(i-1)}\right), \quad i = 1, \ldots, D. \tag{5}$$

Δt is empirically simplified as 1 [50]. Since $\boldsymbol{Q}$ represents the 2D anatomical axial slices, we expect that the evolution of $\boldsymbol{Q}$ should be guided by next-slice image information. Therefore, the velocity field should depend on $\boldsymbol{x}^{(i)}$. Given an INR-based encoder $\mathcal{E}$ that maps $\boldsymbol{x}^{(i)}$ to a latent representation $\boldsymbol{z}^{(i)}$, i.e., $\boldsymbol{z}^{(i)} = \mathcal{E}(\boldsymbol{x}^{(i)})$, the time-dependent velocity field $\boldsymbol{v}^{(i-1)}$ can be written as:

$$\boldsymbol{v}^{(i-1)}\left(\boldsymbol{Q}^{(i-1)}\right) = f\left(\boldsymbol{Q}^{(i-1)}, \boldsymbol{z}^{(i)}\right). \tag{6}$$

Here f refers to the functional mapping for calculating velocity fields, transforming $\boldsymbol{Q}^{(i-1)}$ into $\boldsymbol{Q}^{(i)}$. By incorporating the velocity field $\boldsymbol{v}$ into the image resampling architecture, the implicit decoder $\mathcal{D}$ takes both time-varying representations $\boldsymbol{Q}$ and coordinate embeddings $\boldsymbol{c}$ as inputs to produce a resampled image $\hat{\boldsymbol{r}}$, i.e., $\hat{\boldsymbol{r}} = \mathcal{D}(\boldsymbol{Q}, \boldsymbol{c})$. We train the implicit neural network so that the prediction $\hat{\boldsymbol{r}}$ is close to the ground-truth image $\boldsymbol{r}$.

C. ODE-based Interpolator

The aforementioned time-varying representations $\boldsymbol{Q}$ are generated based on next-slice image latents in an iterative manner. However, since each slice in the volumetric data correlates with both upper and lower slices in the axial direction, single-directional slice progression alone cannot adequately describe the inter-slice dynamics. Therefore, bi-directional ODE constraints are enforced to maintain smoother anatomical variations. The specific formulation is as follows:

$$\boldsymbol{Q}_{\downarrow}^{(i)} = \boldsymbol{Q}_{\downarrow}^{(i-1)} + \Delta t\, f_1\left(\boldsymbol{Q}_{\downarrow}^{(i-1)}, \boldsymbol{z}^{(i)}\right), \quad \boldsymbol{Q}_{\downarrow}^{1} = \boldsymbol{z}^{(1)}, \tag{7}$$

$$\boldsymbol{Q}_{\uparrow}^{(i)} = \boldsymbol{Q}_{\uparrow}^{(i+1)} + \Delta t\, f_2\left(\boldsymbol{Q}_{\uparrow}^{(i+1)}, \boldsymbol{z}^{(i)}\right), \quad \boldsymbol{Q}_{\uparrow}^{D} = \boldsymbol{z}^{(D)}, \tag{8}$$

where $\boldsymbol{Q}_{\downarrow}$ and $\boldsymbol{Q}_{\uparrow}$ correspond to temporal representations in a top-down and bottom-up manner. For the top-down style, the ODE evolution in Eq.(8) requires the initial condition $\boldsymbol{Q}_{\downarrow}^{(1)}$, which is essential for the iterative update. Since $\boldsymbol{Q}_{\downarrow}^{(1)}$ is intended to capture anatomical information of the first slice after the ODE-based interpolator, we directly adopt the latent representation of the first slice $\boldsymbol{z}^{(1)}$ to define it. Analogously, $\boldsymbol{z}^{(D)}$ is set as the initial state of $\boldsymbol{Q}_{\uparrow}^{(D)}$. Through a linear combination of these two elements, we can acquire the time-varying representations $\boldsymbol{Q}^{(i)}$ that satisfy:

$$\boldsymbol{Q}^{(i)} = \underbrace{\frac{\boldsymbol{Q}_{\uparrow}^{(i+1)} + \boldsymbol{Q}_{\downarrow}^{(i-1)}}{2}}_{\textbf{Average}} + \sigma \underbrace{\left[f_1(\boldsymbol{Q}_{\downarrow}^{(i-1)}, \boldsymbol{z}^{(i)}) + f_2(\boldsymbol{Q}_{\uparrow}^{(i+1)}, \boldsymbol{z}^{(i)})\right]}_{\textbf{Instantaneous Velocity Field}}, \tag{9}$$

where σ is a constant value equal to $\Delta t/2$. $\boldsymbol{Q}^{(i)}$ integrates both top-down and bottom-up contextual information, which smooths inter-slice variations and promotes anatomical consistency across slices. More concretely, we decompose $\boldsymbol{Q}^{(i)}$ into two terms. As shown in Eq.(9), the first term is the slice-wise average representation. The second term encodes the influences of instantaneous velocity fields, which warp representations from neighboring slices to slice i, thus yielding more consistent inter-slice evolution.

Next, we will discuss the implementation of functional mapping f_1, f_2 for modeling instantaneous velocity fields between slices. Take f_1 as an example. Velocity fields mainly capture inter-slice differences. Thus, we select $\mathcal{I}_t$ as the target information, as illustrated by the following formulas:

$$\mathcal{I}_t = \left(\boldsymbol{Q}^{(i-1)} - \boldsymbol{V}^{(i)}\right), \tag{10}$$

$$S = \frac{\boldsymbol{Q}^{(i-1)} \cdot \boldsymbol{K}^{(i)}}{\left\|\boldsymbol{Q}^{(i-1)}\right\|_1 \left\|\boldsymbol{K}^{(i)}\right\|_1}, \tag{11}$$

$$f_1\left(\boldsymbol{Q}^{(i-1)}, \boldsymbol{z}^{(i)}\right) = \mathrm{Softmax}\left(1 - S\right)\mathcal{I}_t. \tag{12}$$

Specifically, $\boldsymbol{K}^{(i)}, \boldsymbol{V}^{(i)}$ refers to the key and value transform of $\boldsymbol{z}^{(i)}$. Furthermore, it is significant to determine the attention score within $\mathcal{I}_t$. Indeed, attention should be attenuated for token pairs with high similarity and strengthened for pairs with pronounced differences. Motivated by this point, we adopt the reversed cosine similarity metric. As shown in Eq.(12), $S \in \mathbb{R}^{HW \times HW}$ denotes the pairwise cosine-similarity matrix between $\boldsymbol{Q}^{(i-1)}$ and $\boldsymbol{K}^{(i)}$. We then compute the discrepancy matrix as $1 - S$, where larger values indicate greater dissimilarity between tokens. Notably, this matrix is already properly scaled and can be directly input to the Softmax operation. Finally, the normalized weight is combined with target information $\mathcal{I}_t$ to better model instantaneous velocity fields.

D. Intra-slice Constraint

The ODE-based interpolator can promote segmentation through a unified voxel spacing, as demonstrated by nn-UNet [1], [15]. However, we are encouraged to explore the bi-directional interaction between image resampling and semantic segmentation in this work. Thus, resampling performance is to be improved by introducing additional semantic guidance, which is the key to addressing the ill-posed upsampling problem.

Meanwhile, the proposed interpolator only exploits the inter-slice dynamics of anatomical structures, neglecting the intra-slice contextual correlations. Thus, we leverage a pre-trained vision model to produce high-quality dense representations for 2D slices, which construct a contextual correlation map for the specifically queried location. Specifically, given the dense feature field of a slice $\boldsymbol{Q}^{(i)}$, we compute the correlation between the feature vector at the target point p and all other feature vectors using a cosine-similarity metric, yielding a dense map $\mathcal{A}$ that reflects semantic constraints within the slice. This contextual correlation map enables the identification of a semantically coherent neighborhood around the target point p. The detailed process can be formulated as follows:

$$\mathcal{N}_k(p) = \underset{j \in \mathcal{P}}{\text{Top-k}}\ A(j, p), \tag{13}$$

$$\hat{\boldsymbol{Q}}_p^{(i)} = \sum_{j \in \mathcal{N}_k(p)} \boldsymbol{Q}_j^{(i)} \mathcal{A}(j, p), \tag{14}$$

where $\mathcal{P}$ is the local window centered at point p, $\mathcal{N}_k(p)$ contains points with the top-k highest similarities. By selecting

the top-k most correlated locations $\mathcal{N}_k(p)$, we then perform a weighted aggregation to refine the interpolated features $\boldsymbol{Q}_p^{(i)}$ as enhanced features $\hat{\boldsymbol{Q}}_p^{(i)}$, where the weights $\mathcal{A}(j,p)$ are derived from the normalized similarity scores. By propagating information from semantically consistent neighbors, the proposed strategy promotes class-wise semantic consistency of intra-slice anatomical structures, which is beneficial for the downstream segmentation task.

IV. Experiments

A. Datasets

TopCow 2024 [51]: This public dataset contains 125 brain-vessel MRI image/mask pairs, with voxel spacing ranging from $[0.30, 0.39] \times [0.30, 0.39] \times [0.50, 0.80]$ mm^3. We first hold out 30 cases as the test set for the segmentation task, which reflects the real inference scenario without synthetic spacing perturbation. For each of the 125 cases, we further generate a large-spacing counterpart by downsampling the image along the axial direction with a random scaling factor sampled from $(1.0, 3.0)$. Moderate Gaussian blur and Gaussian noise are also introduced to better simulate the real data acquisition process. In this way, each original image and its downsampled version form a large-spacing/small-spacing pair for the resampling task.

For the segmentation task, the 30 held-out real cases are used only for testing. The remaining 95 real cases and their corresponding downsampled counterparts form 190 image/mask pairs, which are further split into 150 training cases and 40 validation cases. For the resampling task, all 125 large-spacing/small-spacing pairs are split at the case level into 80 training pairs, 15 validation pairs, and 30 testing pairs. The 30 testing pairs correspond to the same held-out cases used for segmentation testing, ensuring that no case-level overlap exists between training and testing. The ground-truth segmentation includes 14 classes, where label 0 denotes the background and labels 1-13 correspond to 13 distinct brain vessels.

BraTS 2020 [52], [53]: BraTS 2020 contains 369 brain MRI scans. Each case has a volume size of $155 \times 240 \times 240$ and an isotropic voxel spacing of $1 \times 1 \times 1$ mm^3. To evaluate the influence of voxel-spacing inconsistency, we adopt the same axial downsampling and Gaussian operations as used for TopCow 2024. Following this protocol, 69 image/mask pairs and 69 corresponding small-spacing/large-spacing pairs are held out for segmentation and resampling inference. For the resampling task, 240 and 60 data pairs are used for training and validation. For the segmentation task, 600 image/mask pairs are split into 480 training pairs and 120 validation pairs. It should be noted that no original case or its downsampled counterpart appears across training and testing splits. The ground-truth segmentation includes four labels: 0 for background, 1 for non-enhancing tumor core, 2 for peritumoral edema, and 4 for GD-enhancing tumor. Due to ambiguous tumor boundaries and highly irregular morphologies, BraTS provides a challenging benchmark for evaluating both resampling fidelity and downstream 3D segmentation performance.

SPIDER [54]: SPIDER is a public lumbar-spine MRI dataset containing 447 scans, including 41 small-spacing cases. Following the same experimental protocol, we construct paired data by resampling high-resolution volumes to coarser axial spacings with random scaling factors twice. 10 image/mask pairs and 10 corresponding small-spacing/large-spacing pairs are held out for segmentation and resampling inference. For the resampling task, 31 data pairs are used for training. For the segmentation task, 93 image/mask pairs are split into 73 training pairs and 20 validation pairs. The ground-truth segmentation contains 20 classes: class 0 denotes the background, classes 1-9 correspond to nine lumbar vertebrae, class 10 denotes the spinal canal, and classes 11-19 correspond to intervertebral discs.

B. Experimental Settings

Implementation Details. For training medical image resampling frameworks, as described in Sec. IV-A, when constructing the training and validation sets, we only apply linear resampling from smaller slice spacing to larger slice spacing to generate paired high-/low-resolution volumes, instead of resampling to an even finer spacing. This choice is motivated by the fact that resampling to smaller spacing may introduce unreliable details and thus lower imaging quality. MedNeXt [7] is selected as the encoder of Consispace, which is trained with the reconstruction loss function. Here we adopt a combination of MSE loss and SSIM loss, and the weights are set as 1.0 and 1.0. All resampling models are trained using the AdamW [55] optimizer with the linear warm-up strategy for the first 25 epochs. The learning rate, batch size and training epoch are equal to $1e$-4, 4, and 500. For the intra-slice semantic constraint, we instantiate the pretrained vision model with DINOv3 [41] to extract dense slice features, and the effect of alternative feature backbones is studied in the ablation experiments.

For training segmentation networks, we use nnU-Net [1], Swin Transformer [6], and MedSAM-2 [24] as baselines to assess the impact of voxel spacing unification. All experiments are trained with the AdamW optimizer, with a warm-up cosine scheduler for the first 50 epochs. The learning rate, batch size, and training epoch are set as $5e$-4, 2, and 1000. The segmentation loss function is the linear combination of Dice loss and Cross entropy loss, with weights equal to 1.0 and 1.0. For data augmentation, we adopt strategies of random rotation between [-15°, 15°], random flipping along the XOZ/YOZ plane, and random intensity scaling between $[0.9, 1.1]$. All experiments are implemented based on Pytorch 2.1.0 and 2 NVIDIA RTX 5090 GPUs.

Evaluation Metrics. For the image resampling task, we select evaluation metrics including Peak Signal-to-Noise Ratio (PSNR) [56], Structural Similarity (SSIM) [57], Normalized Mean Square Error (NMSE) [58], Learned Perceptual Image Patch Similarity (LPIPS) [59]. Compared to PSNR and SSIM, which are traditional pixel-wise reconstruction metrics, LPIPS evaluates perceptual similarities from the perspective of human vision [60], [61]. Therefore, this metric is more feasible and favored in clinical practice. For the segmentation task, the Dice score [62] is adopted as the overlap-based metric, and 95% Hausdorff distance (HD95) [63] is used to measure the performance of boundary delineation.

TABLE I: (a) Quantitative segmentation performance comparisons on different settings of voxel spacing. (b) Comparison with other methods for enhancing the voxel spacing. Two strong segmentation baselines are adopted, including nnU-Net and Swin UNETR. Dice score and HD95 are incorporated as the average metrics for all segmentation classes.

(a) Different settings of voxel spacing							
Model	Setting	SPIDER		TopCow24		BraTS20	
		Dice (%) ↑	HD95 (mm) ↓	Dice (%) ↑	HD95 (mm)↓	Dice (%) ↑	HD95 (mm) ↓
nnU-Net [1]	Original spacing	81.64	14.19	85.59	4.16	85.13	7.06
	Uniform spacing (Linear)	89.13	4.87	88.42	2.72	86.06	6.31
	Uniform spacing (Ours)	**90.56**	**3.48**	**90.83**	**1.88**	**87.20**	**5.39**
Swin UNETR [6]	Original spacing	81.91	15.14	85.72	3.98	84.20	9.84
	Uniform spacing (Linear)	87.28	10.56	87.43	3.19	84.85	8.40
	Uniform spacing (Ours)	**88.32**	**6.84**	**88.76**	**2.56**	**85.72**	**6.97**
(b) Voxel-spacing-oriented methods							
nnU-Net [1]	Synthseg [46]	87.49	6.30	87.17	3.68	85.79	6.05
	HyperSpace [15]	88.79	5.73	89.24	2.37	86.11	5.80
	Ours	**90.56**	**3.48**	**90.83**	**1.88**	**87.20**	**5.39**

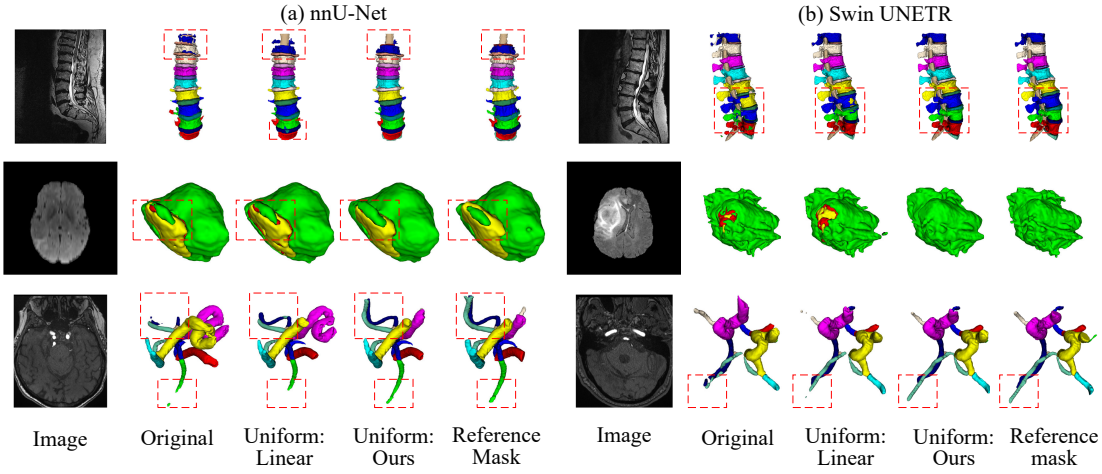

Fig. 3: The performance comparison between linear interpolation and the proposed Consispace on medical image segmentation.

C. Experimental Results

Comparative analysis on the types of voxel spacing. As we mentioned in the Introduction, there are three settings related to the voxel spacing, including:

- Original spacing: All images and masks are trained in a segmentation network, maintaining the original voxel spacing at training and inference time.
- Uniform spacing (Linear): All training images are resampled to the median voxel spacing using linear interpolation, while the corresponding masks are resampled using nearest-neighbor interpolation. The resampled data are then used to train the segmentation network. During inference, images are first resampled to the unified spacing, and model predictions are subsequently mapped back to the original spacing via inverse resampling.
- Uniform spacing (Ours): Compared with the setting of Uniform spacing (Linear), we aim to resample the volumetric images with the proposed Consispace framework in the training and inference stages. The other steps remain the same.

We adopt the classical nnU-Net [1] and Swin UNETR [6] as the segmentation baselines to evaluate quantitative and qualitative performance. As revealed in Table I(a), after unifying the voxel spacing of SPIDER datasets in a vanilla linear way, there is $7.49\% \uparrow$ Dice score and $9.32mm \downarrow$ HD95 based on nnU-Net. This result demonstrates the effectiveness of unifying voxel spacing in the medical image segmentation task, as has been claimed in previous literature [1], [15].

Since linear interpolation is prone to inter-slice non-smoothness, it may result in structural discontinuities. It also leads to distorted boundaries between different anatomies, causing poor semantic consistency within each segmentation class. Thus, we propose Consispace to achieve more effective image resampling. As shown in Table I(a), by replacing linear interpolation with Consispace, there are $1.43\% \uparrow$, $2.41\% \uparrow$, and $1.14\% \uparrow$ Dice score increases on SPIDER, TopCow24 and BraTS20 based on nnU-Net. Besides, we also provide segmentation visualizations among three settings of voxel spacing. As illustrated in Fig.3, resampled images generated by Consispace achieve fine-grained segmentation results, which contain brain vessels with better connectivity, brain tumors with precise boundaries, and vertebrae with intra-class semantic consistency.

Comparison with methods on volumetric image resampling. We investigate the performance analysis of volumetric image

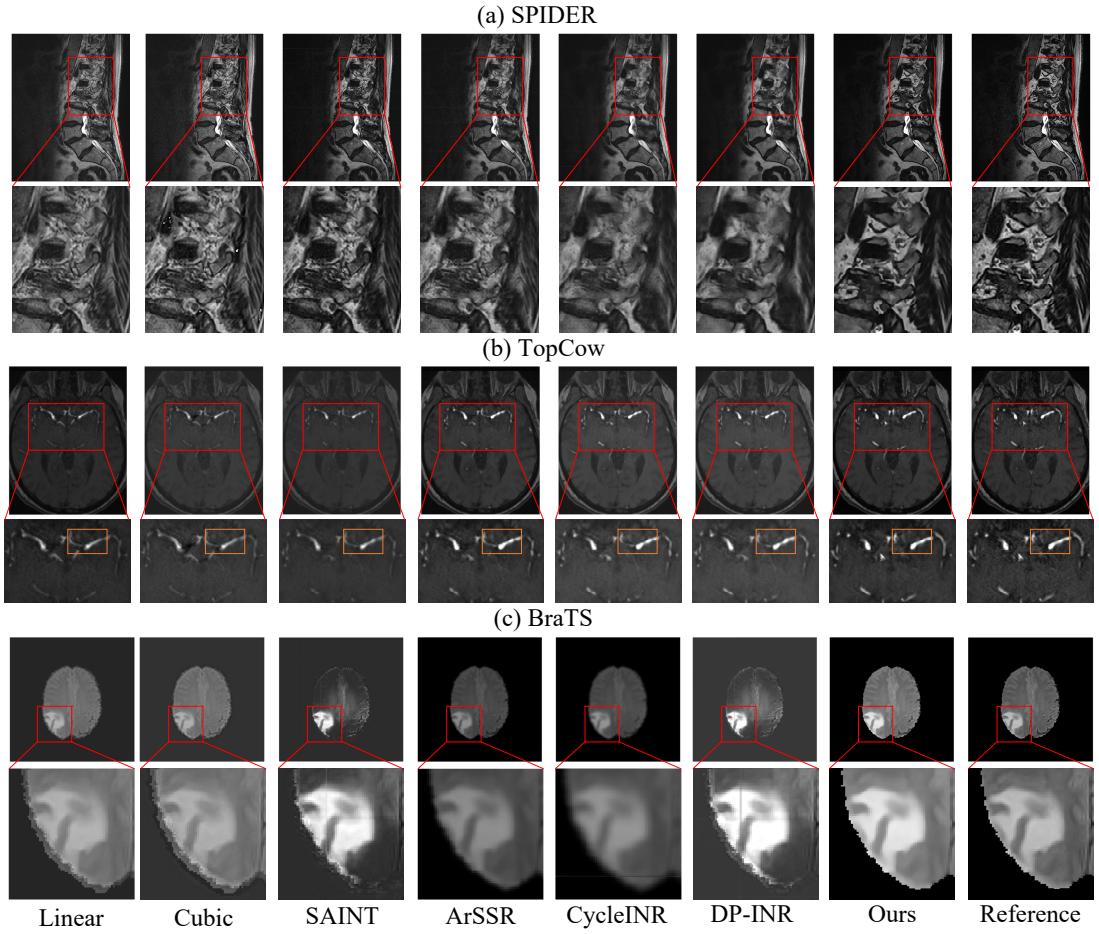

Fig. 4: The image resampling performance comparison between the proposed Consispace and other resampling approaches.

TABLE II: Comparison with other methods on volumetric image resampling when evaluated on SPIDER, TopCow24 and BraTS20. Evaluation metrics include PSNR (dB), SSIM, NMSE, and LPIPS (The best results are highlighted in bold, and the second-best results are underlined).

Method	SPIDER				TopCow24				BraTS20			
	PSNR (dB) ↑	SSIM ↑	NMSE ↓	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓	LPIPS ↓
Linear	23.15	0.563	0.084	2.228	30.89	0.843	0.042	1.429	19.15	0.567	0.333	3.006
Cubic	19.50	0.568	0.156	2.949	25.61	0.745	0.108	1.817	16.95	0.457	0.441	3.694
SAINT [18]	25.55	0.717	0.082	1.912	32.64	0.872	0.034	1.134	23.45	0.876	0.216	2.542
TVSRN [21]	26.87	0.774	0.061	1.746	32.93	0.879	0.037	1.179	24.68	0.889	0.170	2.312
ArSSR [20]	26.32	0.763	0.069	1.714	33.08	0.880	0.030	1.052	24.89	0.903	0.141	2.347
CycleINR [22]	27.20	0.802	0.055	1.651	34.62	0.905	**0.017**	0.891	25.06	0.917	0.101	**1.860**
DP-INR [23]	27.54	0.813	0.052	1.677	34.29	0.897	0.024	0.922	24.74	0.911	0.092	2.129
Consispace	**28.31**	**0.849**	**0.039**	**1.418**	**35.09**	**0.916**	0.019	**0.768**	**25.48**	**0.925**	**0.078**	1.896

resampling. Based on the quantitative results in Table II, our method consistently achieves the best performance on SPIDER, TopCow24, and BraTS20 across four reconstruction metrics. On SPIDER, we obtain 28.31dB PSNR and 0.849 SSIM, while reducing NMSE and LPIPS to 0.039 and 1.418, respectively, outperforming all competing approaches. A similar trend is observed on TopCow24 and BraTS20. Besides, we present qualitative image resampling results. As revealed in Fig.4, Consispace improves both inter-slice and intra-slice fidelity by better preserving fine-grained anatomical textures and boundaries. Specifically, there are fewer generated fake vessels in the upsampled TopCow data, and more accurate vertebral boundaries in SPIDER data.

Comparison with voxel-spacing-oriented approaches. Table I(b) demonstrates that our method yields more significant improvements in downstream segmentation compared with prior voxel-spacing-oriented methods across three datasets. In contrast to Synthseg [46] and Hyperspace [15], we achieve the highest Dice score and the lowest HD95 on testing datasets of SPIDER ($90.56\%/3.48mm$), TopCow24 ($90.83\%/1.88mm$), and BraTS20 ($87.20\%/5.39mm$). Indeed, DINOv3 is incorporated to extract high-quality semantic similarities. These semantic cues are injected by reweighting interpolated features generated by the ODE interpolator, boosting the intra-class semantic consistency within anatomical structures. Thus, the proposed framework effectively bridges the upstream resampling task and the downstream segmentation objective, rather than treating them as two isolated tasks. This contributes to the

TABLE III: Structural ablation on the core components of Consispace, including ODE-based anatomical constraints and semantic consistency constraints derived from DINOv3. We also consider the efficacy of bidirectional anatomical information and reversed attention inside the ODE interpolator.

Settings	SPIDER				TopCow24			
	PSNR (dB) ↑	SSIM ↑	NMSE ↓	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓	LPIPS ↓
Consispace	**28.31**	**0.849**	**0.039**	**1.418**	**35.09**	**0.916**	**0.017**	**0.768**
w/o DINOv3	27.89	0.828	0.048	1.605	34.76	0.905	0.025	0.847
w/o ODE constraints	27.11	0.803	0.059	1.634	34.23	0.891	0.026	0.955
w/o both	26.97	0.796	0.057	1.698	33.10	0.880	0.035	1.119
Unidirectional-ODE	27.97	0.832	0.052	1.613	34.28	0.889	0.029	0.923
Bidirectional-ODE (ours)	**28.31**	**0.849**	**0.039**	**1.418**	**35.09**	**0.916**	**0.017**	**0.768**
Vanilla attention	27.44	0.803	0.056	1.620	34.56	0.903	0.025	0.891
Reversed attention (ours)	**28.31**	**0.849**	**0.039**	**1.418**	**35.09**	**0.916**	**0.017**	**0.768**

TABLE IV: Model efficiency comparison on the resampling task when evaluated on the SPIDER dataset (The inference time refers to the average speed per volume).

Metric	Linear	TVSRN	CycleINR	Ours
Parameters (M)	-	4.05	7.18	2.93
Inference time (s)	0.23	8.32	20.10	14.47
GPU memory (GB)	0.35	7.34	12.09	10.18
FLOPs (G)	4.87	690.50	382.63	237.89

superior downstream segmentation performance of our method compared with other spacing-oriented approaches.

Model efficiency. We further evaluate the computational efficiency of different resampling methods on the SPIDER dataset, as reported in Table IV. As expected, linear interpolation has the lowest computational cost, since it is a non-learnable operation without anatomical or semantic modeling. Among learning-based methods, our method uses the fewest trainable parameters, i.e., 2.93M, compared with 4.05M for TVSRN and 7.18M for CycleINR, where the frozen pretrained vision model weights are excluded from parameter counting. In addition, Consispace requires 237.89G FLOPs, which is substantially lower than TVSRN and CycleINR. These results indicate that Consispace achieves a favorable trade-off between reconstruction performance and computational efficiency.

D. Ablation Study

Structural ablation on Consispace. We first conduct a structural ablation study to evaluate the significance of two key components in Consispace, i.e., the ODE-based anatomical constraints and the intra-slice semantic constraints derived from DINOv3. As reported in Table III, removing either component leads to consistent degradation across all reconstruction metrics on both SPIDER and TopCow24, demonstrating the necessity of both inter-slice anatomical modeling and intra-slice semantic guidance.

Specifically, removing the DINOv3-guided semantic constraint decreases the PSNR from 28.31 dB to 27.89 dB on SPIDER and from 35.09 dB to 34.76 dB on TopCow24, accompanied by increased LPIPS values. This indicates that DINOv3-based semantic correlations provide effective contextual priors for preserving semantically coherent anatomical structures. When the ODE constraints are removed, the performance degradation becomes more pronounced, with the PSNR decreasing to 27.11 dB on SPIDER and 34.23 dB on TopCow24, suggesting the importance of explicitly modeling inter-slice anatomical dynamics. As visualized in Fig. 5, the instantaneous velocity fields enrich the average embeddings with shape-aware information, while semantic reweighting further enhances the reconstructed embeddings with intra-slice semantic correlations. In fact, this work's contribution is not the INR backbone itself, but the integration of inter-slice anatomical continuity and intra-slice semantic consistency into a spacing harmonization framework.

Interaction between image resampling and image segmentation. We further analyze the interaction between volumetric image resampling and downstream medical image segmentation. As discussed in Section IV-C, unifying voxel spacing can reduce the distribution discrepancy caused by heterogeneous physical resolutions and thus improve segmentation robustness. Nevertheless, the effectiveness of spacing unification highly depends on the resampling quality. Conventional linear interpolation relies on intensity-level estimation, which may introduce anatomical discontinuities or blurred boundaries, especially for thin structures and complex anatomical regions.

In contrast, Consispace couples resampling with anatomical and semantic priors. The ODE-based interpolator models continuous inter-slice transitions through bidirectional instantaneous velocity fields, thereby preserving anatomical smoothness along the axial direction. Meanwhile, the DINOv3-guided intra-slice constraint introduces dense semantic correlations into the interpolation process. As shown in Fig. 5, the semantic correlation maps tend to assign higher responses to voxels belonging to the same anatomical or segmentation class. By reweighting interpolated features with these correlations, Consispace enhances class-wise consistency and suppresses less relevant neighboring information. Therefore, Consispace does not treat resampling as an isolated preprocessing step, but establishes a beneficial interaction between resampling and segmentation: semantic priors improve resampling fidelity, while higher-quality resampled images further benefit downstream segmentation.

Application to SAM-based foundation models. To examine the generality of Consispace beyond conventional segmen-

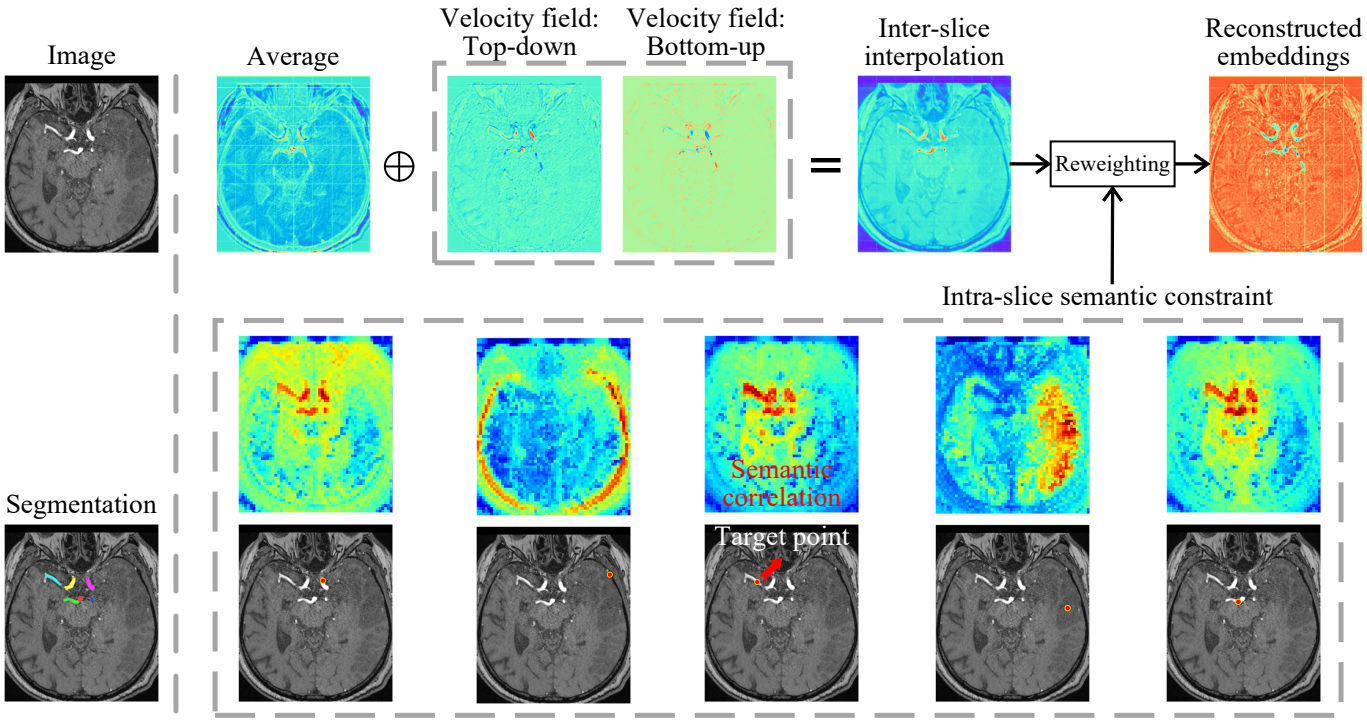

Fig. 5: The detailed feature resampling process in the proposed Consispace.

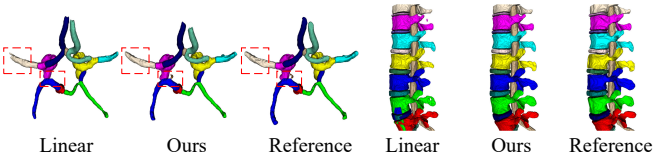

Fig. 6: The segmentation performance promotion when implementing Consispace on the finetuned MedSAM-2 segmentation model.

tation architectures, we further apply it to the SAM-based medical foundation model MedSAM-2 [24]. Specifically, we compare linear resampling and Consispace as preprocessing strategies before fine-tuning MedSAM-2. Quantitative results show that Consispace consistently improves the Dice score, from 84.62% to 86.13% on SPIDER and from 79.33% to 81.89% on TopCow24. These results indicate that the proposed resampling strategy is also beneficial for foundation models adapted to medical image segmentation.

The qualitative comparisons in Fig. 6 further demonstrate that Consispace produces segmentation results with more complete anatomical structures and more accurate boundary delineation. By preserving inter-slice anatomical continuity and intra-slice semantic consistency, Consispace provides more reliable volumetric inputs for MedSAM-2. These findings suggest that Consispace can serve as a general and model-agnostic preprocessing framework for medical image segmentation.

V. Discussion

• ***Backbone selection for feature reweighting.*** DINOv3 [41] is adopted in Consispace to construct intra-slice contextual correlations for reweighting the interpolated feature. To evaluate the influence of the feature backbone, we replace DINOv3 with several representative alternatives, including general-purpose pretrained vision models, i.e., MAE [64] and DINOv2 [65], and medical foundation models, i.e., SAM-Med2D [40] and VISTA3D [66]. As reported in Table VI, DINOv3 achieves the best overall resampling performance on SPIDER, with the highest PSNR of 28.31 dB and SSIM of 0.849. These results suggest that DINOv3 provides more reliable dense representations for identifying semantically correlated neighborhoods within each slice, thereby improving reconstruction fidelity. Compared with MAE and DINOv2, the medical foundation models generally yield better segmentation-related metrics, indicating the benefit of medical-domain priors.

Notably, adopting VISTA3D as the feature selector achieves the highest Dice score of 90.97%, which may be attributed to its strong 3D medical representations and robustness to point-prompt segmentation. However, since VISTA3D operates on volumetric inputs, it requires substantially higher GPU memory during feature extraction. In contrast, DINOv3 achieves the best reconstruction quality and the lowest HD95 of 3.48 mm while maintaining better computational efficiency. Therefore, we adopt DINOv3 as the default backbone for constructing intra-slice contextual maps.

• ***Structural analysis of the ODE-based interpolator.*** We further conduct ablation studies on the internal design of the ODE-based interpolator, including the directionality of ODE evolution and the attention formulation for instantaneous velocity fields. As reported in Table III, replacing the bidirectional ODE with a unidirectional variant consistently degrades reconstruction performance on both datasets. Specifically, the PSNR decreases from 28.31 dB to 27.97 dB on SPIDER and from 35.09 dB to 34.28 dB on TopCow24, with similar trends observed in SSIM, NMSE, and LPIPS. This indicates that

TABLE V: Ablation study on the hyperparameters of the kernel size and value k within the pretrained-vision-model guidance (The best results are highlighted in bold, and the second-best results are underlined).

Kernel size	Value k	SPIDER				TopCow24			
		PSNR (dB) ↑	SSIM ↑	NMSE ↓	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	NMSE ↓	LPIPS ↓
3×3	3	27.75	0.821	0.052	1.610	34.34	0.903	0.027	0.886
3×3	5	27.97	0.830	0.045	1.573	34.71	0.907	0.024	0.859
5×5	3	28.11	0.836	0.040	1.469	34.82	0.909	0.020	0.875
5×5	5	28.31	0.849	**0.039**	**1.418**	35.09	0.916	0.017	0.768
5×5	7	**28.40**	**0.851**	0.040	1.432	**35.17**	**0.919**	**0.016**	**0.752**
5×5	9	28.19	0.841	0.043	1.502	34.98	0.908	0.020	0.806

TABLE VI: Ablation study on the selection of backbones aimed at acquiring the contextual map. Quantitative results are implemented on the SPIDER dataset.

Model	Resampling				Segmentation	
	PSNR ↑	SSIM ↑	NMSE ↓	LPIPS ↓	Dice ↑	HD95 ↓
DINOv2 [65]	27.63	0.814	0.052	1.535	88.87	5.38
MAE [64]	26.51	0.758	0.072	1.795	87.30	6.89
SAM-Med2D [40]	27.96	0.830	0.047	1.618	89.53	4.72
VISTA3D [66]	28.04	0.833	0.041	**1.359**	**90.97**	3.75
DINOv3 [41]	**28.31**	**0.849**	**0.039**	1.418	90.56	**3.48**

single-directional evolution is insufficient to fully characterize axial anatomical variations, whereas bidirectional ODE modeling can integrate contextual information from both neighboring directions and better capture inter-slice transitions.

We also evaluate the attention formulation used for velocity field modeling. Replacing the proposed reversed attention with vanilla attention leads to clear performance drops, with the PSNR decreasing from 28.31 dB to 27.44 dB on SPIDER and from 35.09 dB to 34.56 dB on TopCow24. This suggests that vanilla attention tends to emphasize highly similar or relatively static regions, which is less suitable for modeling velocity fields that aim to capture inter-slice variations. In contrast, reversed attention highlights feature discrepancies between adjacent slices, enabling more accurate estimation of anatomical changes. These results validate the necessity of both bidirectional ODE evolution and reversed attention in the proposed interpolator.

• ***Hyper-parameter analysis of the kernel size and value*** k**.** In this work, the local window is defined as a 5×5 square kernel centered at point p, and the value of k in the proposed Consispace is set to 5. We conduct ablation studies on both the kernel size and the value of k, as reported in Table V. It can be observed that enlarging the local window from 3×3 to 5×5 consistently improves the reconstruction performance on both datasets. For example, when $k = 5$, increasing the kernel size from 3×3 to 5×5 improves the PSNR from 27.97 dB to 28.31 dB on SPIDER, while also yielding better SSIM, NMSE, and LPIPS values. This indicates that a 5×5 local window can capture richer contextual information, thereby providing more reliable local consistency constraints.

Regarding the Top-k operation, increasing k from 3 to 5 brings clear performance gains under the 5×5 kernel setting. Although setting k to 7 achieves the best PSNR and SSIM on both datasets, the improvement over $k = 5$ is marginal. Meanwhile, a larger k inevitably introduces additional computational cost due to the increased number of selected candidates. Moreover, further increasing k to 9 leads to degraded performance, suggesting that incorporating excessive candidates will introduce more irrelevant information. Therefore, we set the kernel size to 5×5 and k to 5, which achieves a favorable trade-off between reconstruction accuracy and computational efficiency.

VI. Conclusion and Future Work

We systematically study the interplay between medical image resampling and segmentation, and propose Consispace, a unified resampling framework that couples upstream reconstruction with downstream segmentation. Consispace incorporates ODE-based anatomical constraints to model inter-slice dynamics, alleviating the limitations of conventional discrete resampling. Moreover, we leverage a pretrained vision model to construct intra-slice correlation maps, injecting class-wise semantic consistency through feature reweighting during resampling. Both intra-slice and inter-slice constraints are integrated into an INR formulation. Extensive experiments demonstrate that Consispace improves reconstruction quality and perceptual fidelity, produces smoother inter-slice anatomical variations, and boosts downstream segmentation performance when used as a preprocessing step. However, this study adopts an anatomy-specific setting rather than a universal resampling-segmentation model. For future work, we will scale up both model capacity and training data diversity to improve generalization to unseen anatomical domains.

References

[1] F. Isensee *et al.*, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021. 1, 2, 3, 4, 5, 6

[2] O. Bernard *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *TMI*, vol. 37, no. 11, pp. 2514–2525, 2018. 1

[3] S. Nikolov *et al.*, "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *arXiv preprint arXiv:1809.04430*, 2018. 1

[4] T. C. Hollon *et al.*, "Near real-time intraoperative brain tumor diagnosis using stimulated raman histology and deep neural networks," *Nature medicine*, vol. 26, no. 1, pp. 52–58, 2020. 1

[5] P. Kickingereder *et al.*, "Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study," *The Lancet Oncology*, vol. 20, no. 5, pp. 728–740, 2019. 1

[6] Y. Tang *et al.*, "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *CVPR*, 2022. 1, 2, 5, 6

[7] S. Roy *et al.*, "Mednext: transformer-driven scaling of convnets for medical image segmentation," in *MICCAI*, 2023. 1, 2, 5

[8] X. You, J. He, J. Yang, and Y. Gu, "Learning with explicit shape priors for medical image segmentation," *IEEE Transactions on Medical Imaging*, 2024. 1

[9] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. B. Ayed, "Boundary loss for highly unbalanced segmentation," *Medical Image Analysis*, vol. 67, p. 101851, 2021. 1

[10] F. Sun, Z. Luo, and S. Li, "Boundary difference over union loss for medical image segmentation," in *MICCAI*. Springer, 2023, pp. 292–301. 1

[11] X. You *et al.*, "Towards boundary confusion for volumetric medical image segmentation," *Medical Image Analysis*, p. 103961, 2026. 1

[12] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical image analysis*, vol. 71, p. 102035, 2021. 1

[13] J. Ma, Z. Yang, S. Kim, B. Chen, M. Baharoon, A. Fallahpour, R. Asakereh, H. Lyu, and B. Wang, "Medsam2: Segment anything in 3d medical images and videos," *arXiv preprint arXiv:2504.03600*, 2025. 1, 2

[14] H. Wang *et al.*, "Sam-med3d: A vision foundation model for general-purpose segmentation on volumetric medical images," *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 1, 2

[15] S. Joutard, M. Pietsch, and R. Prevost, "Hyperspace: Hypernetworks for spacing-adaptive image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 339–349. 1, 2, 3, 4, 6, 7

[16] J. Wasserthal *et al.*, "Totalsegmentator: robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, p. e230024, 2023. 1

[17] X. You *et al.*, "Slord: structural low-rank descriptors for shape consistency in vertebrae segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2025. 2

[18] C. Peng, W.-A. Lin, H. Liao, R. Chellappa, and S. K. Zhou, "Saint: spatially aware interpolation network for medical slice synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7750–7759. 2, 7

[19] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8628–8638. 2

[20] Q. Wu *et al.*, "An arbitrary scale super-resolution approach for 3d mr images via implicit neural representation," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 1004–1015, 2022. 2, 3, 7

[21] P. Yu, H. Zhang, H. Kang, W. Tang, C. W. Arnold, and R. Zhang, "Rplhr-ct dataset and transformer baseline for volumetric super-resolution from ct scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 344–353. 2, 7

[22] W. Fang, Y. Tang, H. Guo, M. Yuan, T. C. Mok, K. Yan, J. Yao, X. Chen, Z. Liu, L. Lu *et al.*, "Cycleinr: cycle implicit neural representation for arbitrary-scale volumetric super-resolution of medical data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 631–11 641. 2, 7

[23] J. Lyu *et al.*, "Diffusion-prior based implicit neural representation for arbitrary-scale cardiac cine mri super-resolution," *Information Fusion*, p. 103510, 2025. 2, 7

[24] J. Zhu, A. Hamdi, Y. Qi, Y. Jin, and J. Wu, "Medical sam 2: Segment medical images as video via segment anything model 2," *arXiv preprint arXiv:2408.00874*, 2024. 2, 5, 9

[25] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015. 2

[26] Ö. Çiçek *et al.*, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *MICCAI*, 2016, pp. 424–432. 2

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 2

[28] J. M. J. Valanarasu *et al.*, "Unext: Mlp-based rapid medical image segmentation network," in *MICCAI*. Springer, 2022. 2

[29] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021. 2

[30] J. Chen, J. Mei *et al.*, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024. 2

[31] H.-Y. Zhou, J. Guo, Y. Zhang, X. Han, L. Yu, L. Wang, and Y. Yu, "nnformer: Volumetric medical image segmentation via a 3d transformer," *IEEE transactions on image processing*, vol. 32, pp. 4036–4045, 2023. 2

[32] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021. 2

[33] Z. Cheng, J. Guo, J. Zhang, L. Qi, L. Zhou, Y. Shi, and Y. Gao, "Mamba-sea: A mamba-based framework with global-to-local sequence augmentation for generalizable medical image segmentation," *IEEE Transactions on Medical Imaging*, 2025. 2

[34] Z. Zhang *et al.*, "Switch-umamba: Dynamic scanning vision mamba unet for medical image segmentation," *Medical Image Analysis*, p. 103792, 2025. 2

[35] Z. Xing *et al.*, "Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation," *IEEE Transactions on Medical Imaging*, 2025. 2

[36] J. Liu, H. Yang, H.-Y. Zhou, L. Yu, Y. Liang, Y. Yu, S. Zhang, H. Zheng, and S. Wang, "Swin-umamba†: Adapting mamba-based vision foundation models for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 44, no. 10, pp. 3898–3908, 2024. 2

[37] J. Ma, F. Li, and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024. 2

[38] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026. 2

[39] J. Wu *et al.*, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *Medical image analysis*, vol. 102, p. 103547, 2025. 2

[40] J. Cheng *et al.*, "Sam-med2d," *arXiv preprint arXiv:2308.16184*, 2023. 2, 9, 10

[41] O. Siméoni, H. V. Vo *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025. 2, 5, 9, 10

[42] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. 2

[43] W. Zhan, M. Lin, C.-W. Lin, and R. Ji, "Anysr: Realizing image super-resolution as any-scale, any-resource," *IEEE transactions on image processing*, vol. 33, pp. 6564–6578, 2024. 3

[44] Z. Zhang, L. Zhang, Y. Cheng, Z. Wang, F. Wang, H. Zhang, Y. Yang, Y. Wu, J. Huang, A. I. Aviles-Rivero, Z. Gao, G. Yang, and P. J. Lally, "From coarse to continuous: Progressive refinement implicit neural representation for motion-robust anisotropic mri reconstruction," *IEEE transactions on image processing*, vol. 35, pp. 3550–3565, 2026. 3

[45] C.-H. Qiu, X.-S. Zhang, and Y.-J. Li, "Joanet: An integrated joint optimization architecture making medical image segmentation really helped by super-resolution pre-processing," *IEEE transactions on image processing*, vol. 34, pp. 7008–7023, 2025. 3

[46] B. Billot *et al.*, "Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining," *Medical image analysis*, vol. 86, p. 102789, 2023. 3, 6, 7

[47] E. Kondrateva, P. Druzhinina, and A. Kurmukov, "Do we really need all these preprocessing steps in brain mri segmentation?" in *Medical Imaging with Deep Learning*, 2022. 3

[48] S. Seoni *et al.*, "All you need is data preparation: A systematic review of image harmonization techniques in multi-center/device studies for medical support systems," *Computer Methods and Programs in Biomedicine*, vol. 250, p. 108200, 2024. 3

[49] M. F. Beg *et al.*, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International journal of computer vision*, vol. 61, no. 2, pp. 139–157, 2005. 3

[50] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022. 4

[51] K. Yang *et al.*, "Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra," *ArXiv*, pp. arXiv–2312, 2025. 5

[52] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014. 5

[53] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018. 5

[54] J. W. van der Graaf, M. L. van Hooff, C. F. Buckens, M. Rutten, J. L. van Susante, R. J. Kroeze, M. de Kleuver, B. van Ginneken, and N. Lessmann, "Lumbar spine segmentation in mr images: a dataset and a public benchmark," *Scientific Data*, vol. 11, no. 1, p. 264, 2024. 5

[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017. 5

[56] A. Hore and D. Ziou, "Image quality metrics: Psnr vs. ssim," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 2366–2369. 5

[57] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 5

[58] J. Kim, H. Yoon, G. Park, K. Kim, and E. Yang, "Data-efficient unsupervised interpolation without any intermediate frame for 4d medical images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 353–11 364. 5

[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595. 5

[60] S. Jain, D. Watson, E. Tabellion, B. Poole, J. Kontkanen *et al.*, "Video interpolation with diffusion models," in *CVPR*, 2024, pp. 7341–7351. 5

[61] X. You *et al.*, "Fb-diff: Fourier basis-guided diffusion for temporal interpolation of 4d medical imaging," in *ICCV*, 2025, pp. 28 010–28 020. 5

[62] F. Milletari *et al.*, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571. 5

[63] D. Karimi and S. E. Salcudean, "Reducing the hausdorff distance in medical image segmentation with convolutional neural networks," *TMI*, vol. 39, no. 2, pp. 499–513, 2019. 5

[64] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009. 9, 10

[65] M. Oquab, T. Darcet *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023. 9, 10

[66] Y. He *et al.*, "Vista3d: A unified segmentation foundation model for 3d medical imaging," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 863–20 873. 9, 10